

# Cultural Representation in American Art Museums

---

Sarah Moore & Molly Stark

[Github](#)

## Background

In our experience world art courses, exhibits and information has been limited to the North America and Europe with little representation for Asia, Africa, South America, and Oceania. This is disappointing for those wishing to learn about art around the world. The Metropolitan Museum of Art (MET) and The Art Institute of Chicago are two of the biggest art museums in the United States. The MET has 1.5 million pieces in their collection with over 450,000 of those catalogued in their github. The Art Institute of Chicago has over 300,000 pieces with 130,000 catalogued in their github dataset. We wanted to take a look through their catalogued collections on github to see how well different countries were represented.

## Motivation

Our hypothesis is that many U.S. art museums, which aim to present and educate the public on global art, may be lacking in cultural diversity. A limited study by Topaz et al. (2019) found that of the approximately 10,000 identifiable artists across 18 U.S. museums, 75.7% were white men. This suggests that the cultural representation in these museums is heavily skewed toward North American and European artists. However, it is worth noting that some institutions, such as the Contemporary Arts Museum Chicago, have made efforts to diversify their collections, beginning as early as 1990 (Sweeney et al., 2018). Our mission is to see what areas of cultural representation are lacking in these two museums' collections, so that their acquisition teams could use these findings to diversify and enhance their collections.

## Objectives

1. Gather data on the country of origin for each piece of artwork in the two datasets.
2. Explore what countries and continents have the most and least representation in each museum, separately and combined.
3. Compare population and land mass of each continent to the representation it receives in the art museums.

# Data Sources

| Source Name   | Description  | Important Variables  | Format & Size                | Access   |
|---|--|--|------------------------------|--|
| <b>The Metropolitan Museum of Art Open Access CSV</b> | A partial inventory of the artwork owned by the Metropolitan Museum of Art located in New York City, NY. Last updated June 2023. | <b>Country:</b> country artwork was made in<br><b>Culture:</b> what culture the piece was created in<br><b>Artist Nationality:</b> country artist is from<br><b>Region:</b> region of the world work was created<br><b>Acquisition Date:</b> year that the museum acquired the artwork | CSV file, 303 MB             | Downloadable directly from the MET Museum's GitHub page.<br><br><a href="https://github.com/metmuseum/openaccess">https://github.com/metmuseum/openaccess</a>  |
| <b>The Art Institute of Chicago Public API Data</b>   | A partial inventory of the artwork owned by the Art Institute of Chicago, located in Chicago, IL. Last updated February 2025.    | <b>Title:</b> name of the artwork<br><b>Artist Title:</b> artist's name<br><b>Place of origin:</b> location as address, city, or country art was created<br><b>Acquisition Date:</b> year that the museum acquired the artwork   | 132,785 JSON Files, 688.5 MB | The folder of JSON files containing this data is downloadable from the AIC's GitHub page.<br><br><a href="https://github.com/art-institute-of-chicago/api-data/tree/master/json/artworks">https://github.com/art-institute-of-chicago/api-data/tree/master/json/artworks</a> |



**Goal:** Extract country data for each piece in our sample of the collection. Match each country with its continent.

**Pre-processing:** Due to the large file size, the original CSV was uploaded to Excel's Data Query. Then a column was added and each row assigned a random value between 0-1. The dataset was then sorted in ascending order and the first 150,000 rows were selected to create a new smaller random sample and saved as a CSV file.

## Data Manipulation Method

### The Metropolitan Museum of Art (MET)

1. **Combine Columns:** We used the `reduce` and `combine_first` functions to assign rankings to key columns that potentially contained country data for each artwork. The columns were ranked based on their likelihood of containing country-related information, in the following order: country, culture, artist nationality, region, subregion, and tags. A new column was then created, storing the highest-ranked column with a value for the country. (MET\_data\_cleaning.ipynb, cells 6-9)
2. **Country Dictionary:** A data frame containing country names, alternate names, and demonyms was loaded and converted into a dictionary. This dictionary will be used to map the values in the newly created country column to standardized country names later. (MET\_data\_cleaning.ipynb, cells 13-23)
3. **Regex Application:** We applied regular expressions (regex) to clean the new country column by removing white spaces, extraneous punctuation, and interfering words. Some complex cases arose, such as "South China," which had to be manually added to the dictionary as "China". If words like "south" were removed, it could prevent proper mapping for entries like "South Korea." (MET\_data\_cleaning.ipynb, cell 25)
4. **Mapping:** After combining the columns and applying the regex clean-up, the country dictionary was mapped to the newly created column, replacing the original values with a standardized country name where possible. A separate column was created to preserve the original values, allowing for verification of mapping accuracy. (MET\_data\_cleaning.ipynb, cells 27-29)
5. **Manual Investigation:** We manually reviewed entries that did not match any dictionary item to determine if they could be linked to a specific country. Some cultures were too broad to be associated with a single country, so these were left out. Similarly, any entries without a discernible country were left blank. Given the dataset's large size, we were confident that we still had a good sample even with 8.6% missing values. (MET\_data\_cleaning.ipynb, cells 30-35)
6. **PyCountry-Convert:** We utilized the PyCountry-Convert package to convert the country names into continent names, which were stored in a new column. This allowed us to explore the representation of different continents within the dataset. (MET\_data\_cleaning.ipynb, cells 36-40)

# ART INSTITUTE CHICAGO

**Goal:** Extract country data for each piece in our sample of the collection. Match each country with its continent.

**Pre-processing:** After downloading the 'artworks' folder containing all of the relevant JSON files from the Art Institute of Chicago's GitHub repository, a CSV file containing all of the raw data in a single dataframe was created using the glob module and reading through each file

## Data Manipulation Method

### Art Institute of Chicago (AIC)

1. **Dropping rows with no location data:** We filtered the data frame so that no rows with a null value in the 'place\_of\_origin' column would remain in the data set. We found that only 7% percent of rows were dropped, which left plenty of usable data for our analysis. (AIC\_data\_cleaning.ipynb, cells 9-10)
2. **Categorizing location data:** Once we were only left with data that had a 'place\_of\_origin' value, we created 2 lists of tuples, each containing an artwork ID and its place of origin. One list contained tuples of artworks where their place of origin matched a list of all country names. The other list contained tuples of all the artworks whose place of origin did not match up with a value in our list of countries, allowing us to see what rows needed to be converted. (AIC\_data\_cleaning.ipynb, cells 12-15)
3. **Converting non-country location to a country:** In order to convert all of the data from the rows that did not have location information that aligned with our list of countries, we used the geopy package that allowed us to make API calls to Nominatim. This allowed us to search the OpenStreetMap API and get the resulting address output for each of our non-country locations. We ran into many problems with this API, namely that it only allowed us to make one call every second and would time out often. To get around this, we implemented a try-except function that would cause the program to sleep for a few seconds if it was going to time out, allowing us to get through all the data. (AIC\_data\_cleaning.ipynb, cells 17-18)
4. **Extracting country from address:** The geopy package returned a string of an address for each location we fed it. In order to extract the country name from the address, we used regex to only keep any part of the address that came after the final comma in the string. (AIC\_data\_cleaning.ipynb, cells 21-22)
5. **Cross-checking new country data with country list:** We then checked to see if the new country data all aligned with the list of countries we used. When we found that a few country names did not match the names we were using, we created a dictionary of those countries and the names we were using for them in order to replace those names. (AIC\_data\_cleaning.ipynb, cells 24-30)
6. **Pycountry-Convert:** We used the pycountry-convert package to create a new column containing the continent that each country corresponds to into a new column. (AIC\_data\_cleaning.ipynb, cells 32-35)

# Combining Datasets

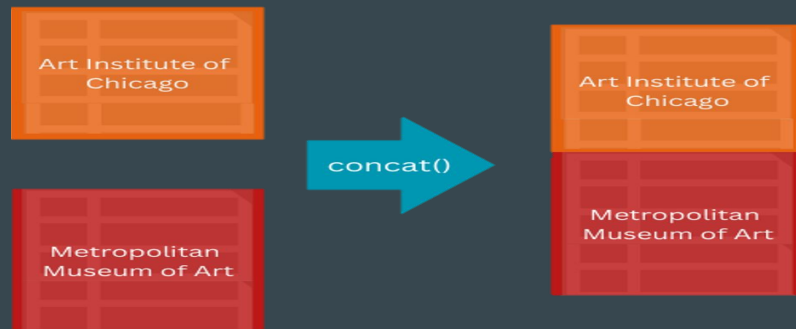
After finishing processing the data and creating clean data sets for use in our analysis, there were still a few steps we needed to take to prepare both data sets to be combined:

1. **Comparing the column names of both data sets:** By printing out the list of column names from both processed data sets, we saw which columns were shared between both sets and determine which of those would be relevant to our analysis. The columns we decided to keep are: 'title', 'department', 'accession\_year', 'country', and 'continent'. (visualizations\_final.ipynb, cells 7-8)
2. **Renaming columns to match between both data sets:** Some of the column names in the AIC processed data set did not align with their corresponding columns in the MET processed data set. By renaming those columns to match the names in the MET data, we ensured that we could concatenate both data sets together and have them properly match up. (visualizations\_final.ipynb, cell 10)
3. **Dropping columns that are irrelevant or are not present in both data sets:** The next step we took before combining the processed data was to drop any columns that were still present in the processed data sets that we did not want to appear in the final combined data set. (visualizations\_final.ipynb, cell 11)
4. **Adding a 'location' column to both data sets:** As a final step before combining the data, we added a 'location' column to both data sets, with a value of 'MET' for any rows in the MET data set, and 'AIC' for rows corresponding to AIC artworks. This allows us to easily keep track of which data set each row originated from in the combined data set. (visualizations\_final.ipynb, cell 12)

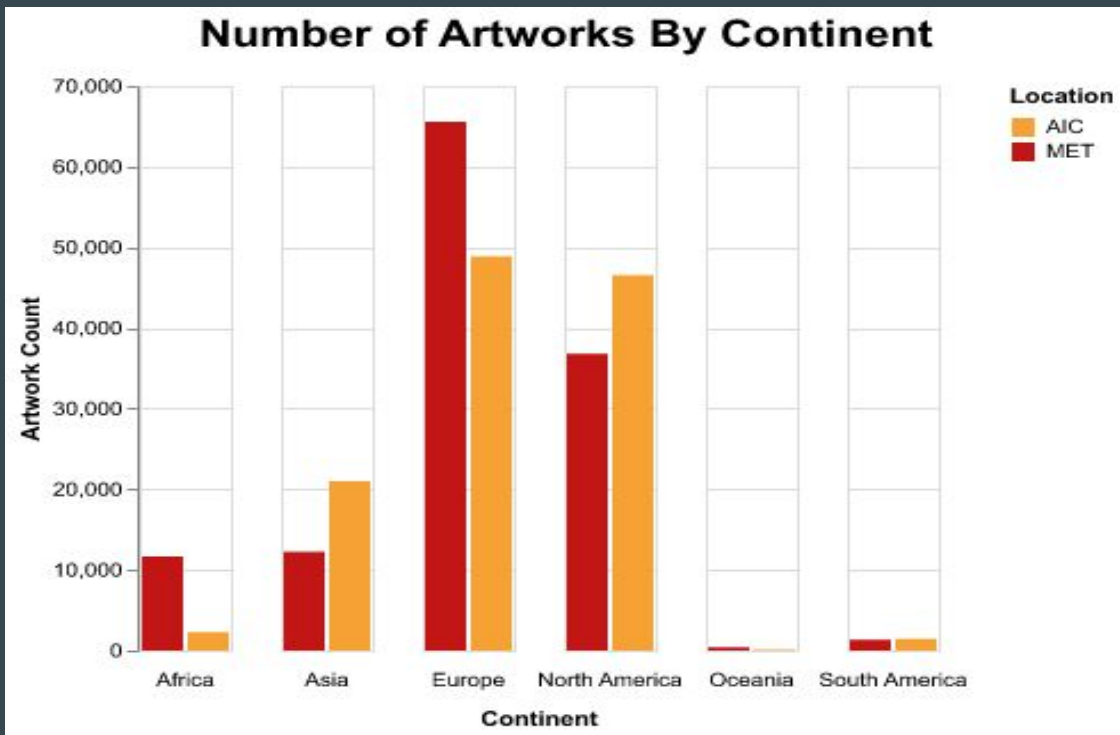
## Final Step: Combining via Concatenation

To create the combined data set, we used the Pandas `.concat()` function to concatenate the data sets together by essentially stacking one on top of the other, while keeping them aligned by their columns. This resulted in a single data set that combined all rows from both museums' processed data sets.

After creating this massive dataframe, we developed 2 other combined data sets to help us create relevant visualizations. Both of these data sets were created by grouping the data by both the museum location column, and the country name / continent name, and then summing the number of rows in each group. This gave us the counts of artworks from each museum, that come from each country or continent, allowing us to more easily compare the places that have the most representation between both museums.



# Continent Bar Chart Visualization & Analysis



**Analysis:** We used the `count()` function to count how many times each continent appeared in the continent column after grouping the data frame by continent. Then created a dataframe with the continent and their counts.

**North America/Europe:** The bar chart reveals a clear imbalance in the distribution of artworks between continents in the Art Institute of Chicago and The MET, with the distribution heavily skewed towards North America and Europe.

**Africa/Asia:** The next highest level of representation seen in both museums is in Africa and Asia. The Art Institute of Chicago shows a fairly equal focus on Africa and Asia, while The MET has a significantly larger share of artworks from Asia compared to Africa.

**South America/Oceania:** Both museums display a notable lack of works from Oceania and South America. While Oceania's population is less than 10% of North America's, explaining its limited presence, South America's population of about 443 million—compared to North America's 608 million—suggests it should have a more substantial presence in the collections.

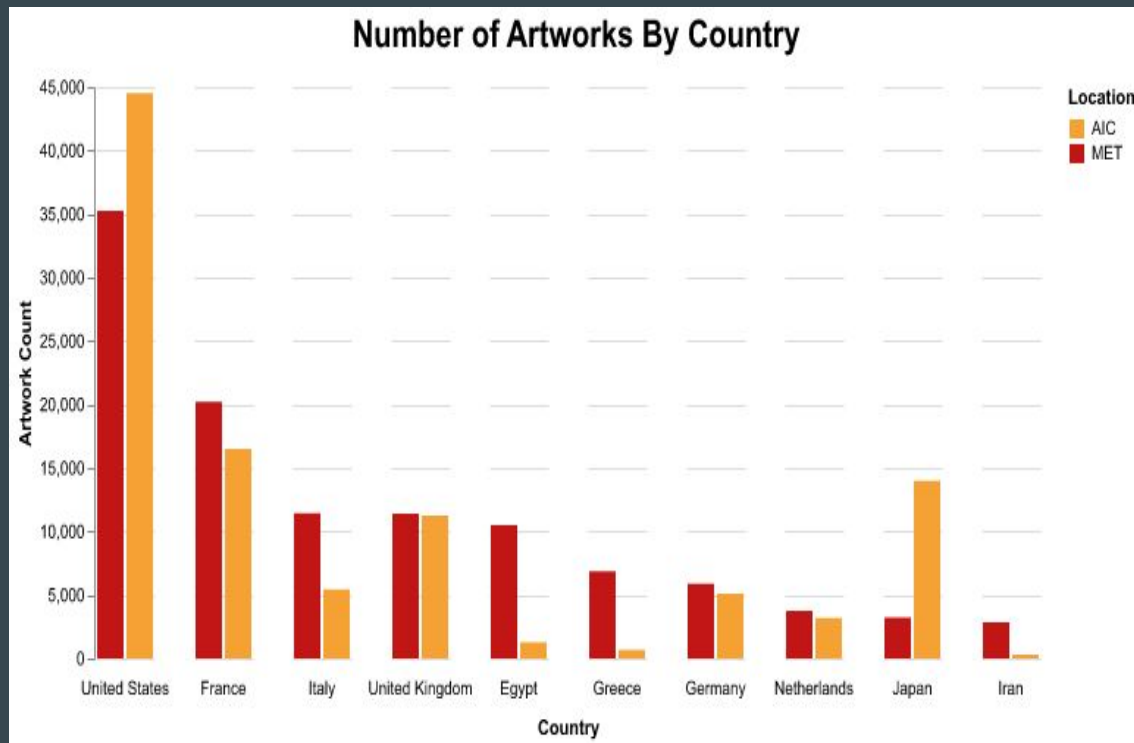
# Country Bar Chart Visualization & Analysis

**Analysis:** We used the `count()` function to count how many times each country appeared after grouping the data frame by country. Then created a dataframe with the country and their counts.

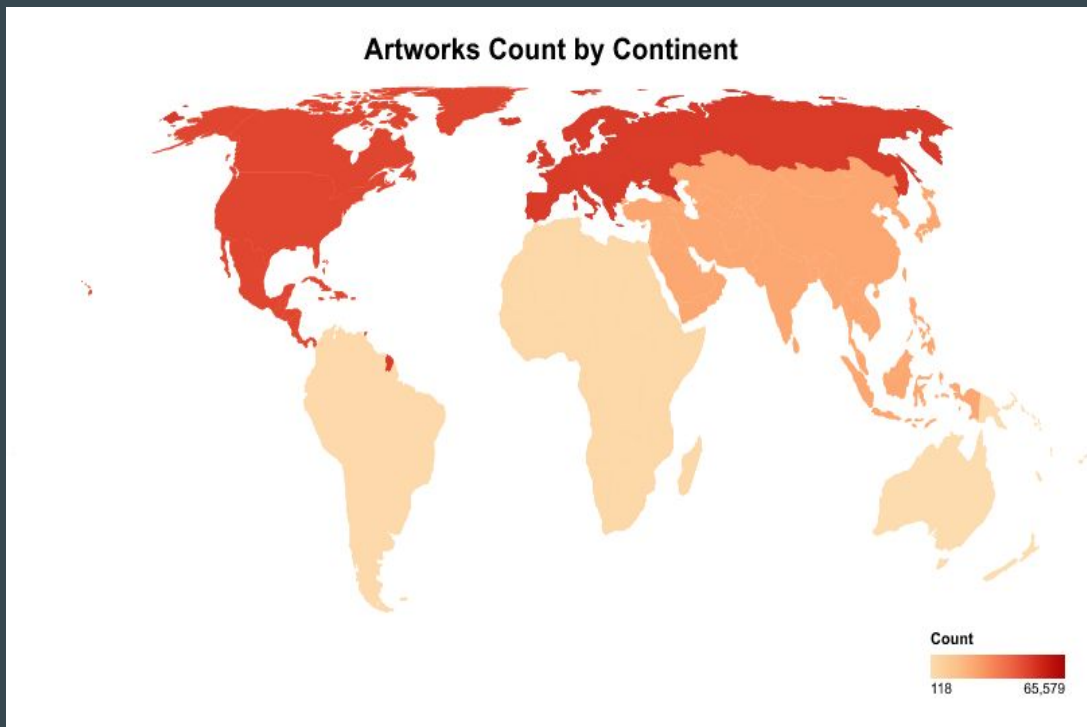
**United States:** Both datasets show a pronounced skew toward the United States in their collections. While the Art Institute of Chicago's concentration of U.S. artworks is particularly high, The MET also demonstrates a notable U.S. presence. This trend is understandable, as acquiring art within one's own country is far easier than sourcing pieces from abroad where transportation is more complex and customs regulations must be followed.

**European Countries:** France holds the second-highest representation in both museums, with its number of artworks approaching half of the U.S. total. The other six European countries represented show a more modest share, but their numbers remain relatively consistent with each other.

**Outliers:** Both datasets feature one country outside of the U.S. and Europe that stands out with a significant representation. The MET includes a substantial amount of Egyptian art, while the Art Institute of Chicago has a notable collection from Japan.



# Combined Dataset Continent Choropleth Chart Visualization & Analysis



**Choropleth:** We selected a choropleth to illustrate the representation of continents in our combined dataset because it offers a visually compelling way to highlight the disparity in representation. We chose a choropleth over a bar chart, so the similarities/differences in land size could be evaluated alongside the distribution. Concatenating the datasets after standardizing the column names allowed us to create a combined museum choropleth easily and effectively.

**Combined representation:** The choropleth clearly illustrates that the combined representation in North America and Europe far outweighs that of the other continents, despite their similar land sizes. Asia follows with over 30,000 artworks across the two museums. In contrast, Africa, South America, and Oceania are significantly underrepresented, with fewer than 30,000 artworks combined. When comparing these continents by landmass, South America and Africa, in particular, should have a representation closer to that of Asia and Europe. Oceania is the only continent where the representation is expected to be lower due to its size.

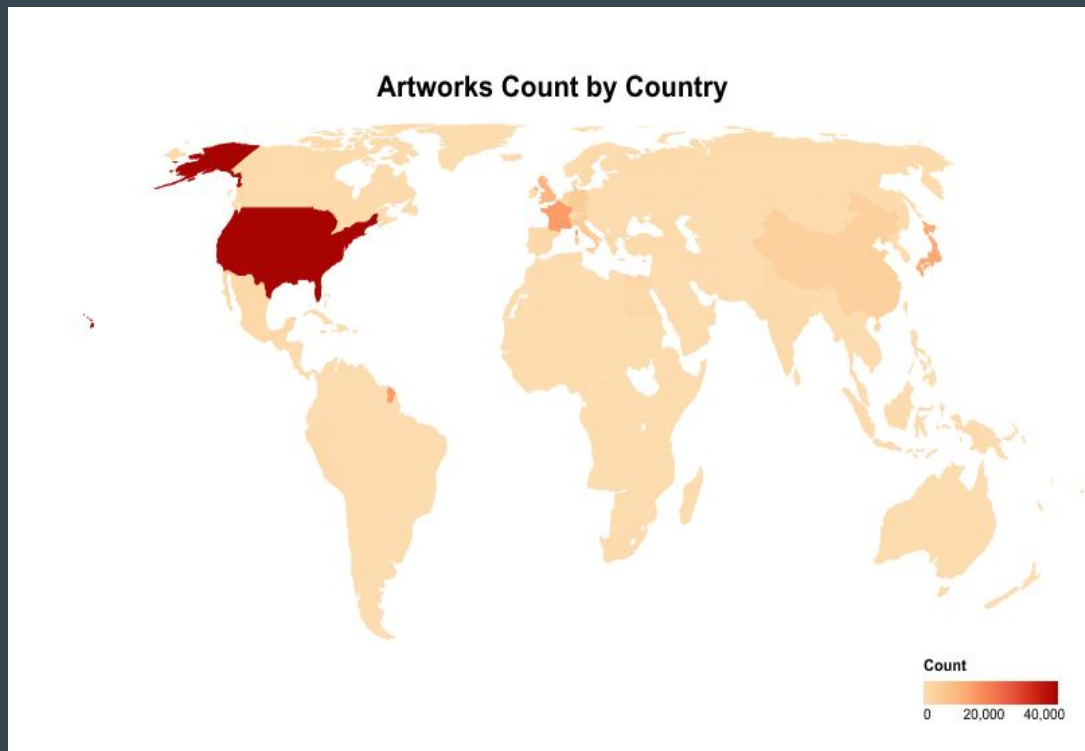
# Combined Dataset Continent Choropleth Chart Visualization & Analysis

**United States:** The United States, predictably, is the darkest red and overwhelmingly the most represented country in the combined data.

**France:** What makes this choropleth interesting is how dark France is compared to the rest of Europe. Would Europe as a whole have been as well represented if France's representation looked more like the rest of Europe?

**Japan & China:** Although Japan is a small country, it is darker than many of the other countries in this choropleth. Even China, which is many times larger than Japan and part of the same continent is slightly lighter.

**African Countries:** It is interesting to note that going from the continent map to the country map, Africa looks like a Swiss cheese version of itself due to some countries having no representation in either museum. This shows how seriously lacking African representation is in these museums, even with Egypt being a larger contributor at the MET.



## Conclusion

Our analysis reveals a significant imbalance in the representation of artworks across continents at the Art Institute of Chicago and The MET. North America and Europe dominate the collections, with a strong focus on the United States and European countries. This is predictable, as acquiring art within a museum's own country is easier.

Asia follows with notable representation, especially in The MET, which features slightly more Asian art than African art. In contrast, Oceania and South America are significantly underrepresented. Oceania's limited population may explain its low presence, while South America's population suggests it should have a larger share. Africa is similarly underrepresented, despite large exhibits on specific countries like Egypt.

The choropleth highlights some interesting country-level trends. France stands out within Europe, showing a higher concentration of artworks than most other European nations. Other European countries show a more modest but consistent presence. Meanwhile, Japan, despite being smaller in size, appears more prominently in the data than larger countries in Asia, such as China.

The African continent's underrepresentation is especially apparent, with some countries, like Namibia, showing no presence in either museum. Egypt having a bigger presence gives a limited scope for museum goers wanting to see a comprehensive collection of African art.

## Further Work

**Accession Dates:** This work could be taken further by cross referencing the accession dates with where different artworks have come from. This would show us trends in accession of artwork and could be linked back to social shifts or art directors during a specific time period.

**Departments:** Cultural diversity of artwork could be explored at the departmental level. Do some departments have more diversity than others? Are some departments less diverse due to material availability across cultures?

**Increase scope:** To be sure this is an accurate representation of American museums, further steps should include incorporating data from more art museums. The main difficulty would be finding data as not all museums have easily accessible data for their collections or the data is difficult to sift through.

## Limitations

Our analysis was hindered most by the size of our files. We had to take a random sample of the MET's data because the original file size was too big to be uploaded to our notebook. Both museums only provide a sample of their collections which are at least twice as big as their datasets. This could have skewed our results, especially if the pieces not included in their dataset are in exhibits focused on certain countries or regions.

# Statement of Work

**Molly Stark:** Revised project proposal  
Filled out and submitted status updates  
Cleaned and manipulated AIC data  
Created exploratory visualizations  
Created and polished final visualizations  
Standardized notebooks and Github repository  
Wrote out data sources, AIC data manipulation, and combining slides  
Revised report

**Sarah Moore:** Came up with project idea and found data sources  
Wrote out first draft of proposal  
Found country dataset  
Cleaned and manipulated MET data  
Created exploratory bar chart visualizations  
Wrote out MET data manipulation slide and remaining slides  
Put together report and formatted slide deck

**Collaboration:** Collaboration went really well. We kept in regular contact with each other via zoom and Slack. We also let each other take the lead where we felt we had expertise. Collaboration could have been improved by meeting more regularly in the beginning, so we could have gotten our data sources solidified and started cleaning data sooner.

# References

- art-institute-of-chicago. (2020).  
*GitHub-art-institute-of-chicago/api-data*. GitHub.  
<https://github.com/art-institute-of-chicago/api-data>
- GeoPy. (n.d.). *Welcome to GeoPy's documentation! — GeoPy 1.21.0 documentation*. Geopy.readthedocs.io.  
<https://geopy.readthedocs.io/en/stable/>
- Python Software Foundation. (n.d.). glob — Unix style pathname pattern expansion — Python 3.8.3rc1 documentation. Docs.python.org.  
<https://docs.python.org/3/library/glob.html>
- Macfarquhar, C. (1994). Encyclopedia Britannica. In *Encyclopædia Britannica*. <https://www.britannica.com/>
- metmuseum. (2016). *GitHub - metmuseum/openaccess: The Metropolitan Museum of Art's Open Access Initiative*. GitHub.  
<https://github.com/metmuseum/openaccess>
- mledoze. (2025). *countries/dist at master · mledoze/countries*. GitHub.  
<https://github.com/mledoze/countries/tree/master/dist>
- PyPi. (2018, February 18). *pycountry-convert: Extension of Python package pycountry providing conversion functions*. PyPI.  
<https://pypi.org/project/pycountry-convert/>
- Python. (2009). *re — Regular expression operations — Python 3.7.2 documentation*. Python.org.  
<https://docs.python.org/3/library/re.html>
- Sweeney, L., & Daniel, K. (2018, October 10). *At Fifty, Remodeling for Equity - Ithaka S+R*. Ithaka S+R.  
<https://sr.ithaka.org/publications/at-fifty-remodeling-for-equity/>