Project 2: Inside the Met

A closer look at the treasures at The Metropolitan Museum of Art through a data science perspective

Team: Stephanie He, George Jiao, and Tiantian Zhao
GitHub repository address: https://github.com/UC-Berkeley-I-School/Project2_He_Jiao_Zhao.git

Overview

The Metropolitan Museum of Art of New York City, known as "the Met", is the largest art museum in the United States. With over 2 million artworks as permanent collection, it offers any visitor almost any common form of visual arts, ranging from ancient Egypt art to modern art, to experience and enjoy. However, perhaps even the most efficient and experienced visitors will not be able quick browsing through a meaningful amount of collection within a short period of time.

Thanks to the Met Museum openaccess, nearly a quarter of (~470k) of artworks have been indexed and its key specifics categorized into a dataset (the "dataset"). Based on this dataset, we intend to present an overview, from top-down perspective, on what's included in the Met, through a thorough analysis on the dataset, and visualization of key findings.

Source Data

Dataset will come from an opensource csv file laying out specifics of 470k pieces of artworks from Github account of the Met < https://github.com/metmuseum/openaccess>

The csv file has 475,125 rows and >50 columns.

Initial Exploration and Data Preparation

The initial exploration includes data completeness and cleanness. From data completeness perspective, there are 53 variables in this dataset, 10 variables have no missing values, 21 variables have less than 50% missing values, and 22 variables have more than 75% missing values.
The key takeaway from this exploration is nearly half of the variables are in a relatively high quality in terms of completeness. Team decides to focus on this half as possible in order to achieve a quality analysis. Those variables include but not limited to *Object Number, Country, Period, Dynasty, Reign, Artist Display Name, Artist Nationality, is Public Domain, Object Name, Accession Year, Object Begin Date and Object End Date etc.*

After deciding the variables to be analyzed, here are few basic data preparation steps done.

- Read in data and rename variables per coding good practices, such as replace space with underscore and/or make them all lower case.
- Identify and process missing values.
- Identify and process special characters in numeric variables.
- Identify and process outliers and dirty data.
- Subset data for each graph/analysis

One specific challenge team encountered is unexpected data format. For example, in the variable name – AccessionYear, there are only 2 special cases where the format is not a year but a full date with slash, those has been causing unexpected bumpers to team. This is a great lesson-learned that, when data is large enough there can be special cases hidden almost everywhere, need to be prepared for using an effective code to locate and process them.

Exploratory Questions

Based on the initial data exploration results, from feasibility perspective, the team decided to focus on two major questions with their sub questions as follows:

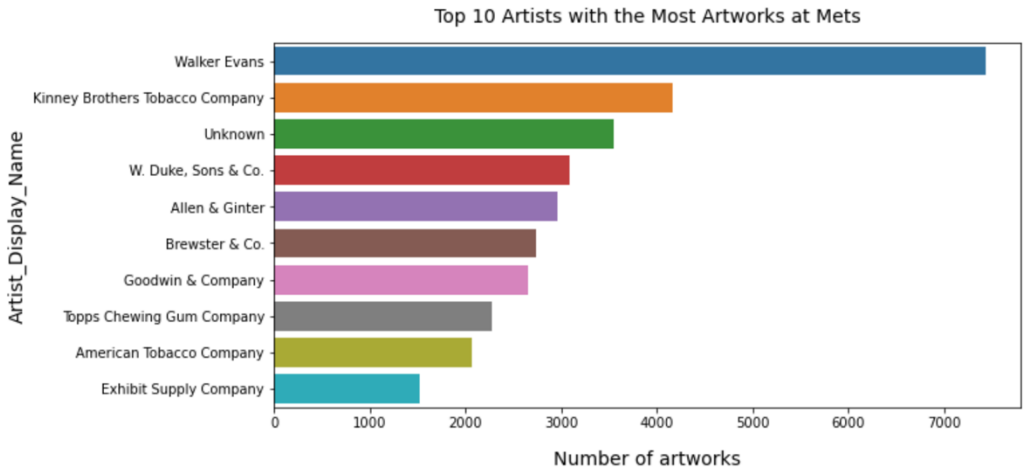
1. **Is the Met the same with how it appears to be when we visit?**
    - 1.1. Who is the most collected artist? What is his/her work?
    - 1.2. Visitors are easily attracted by painting. So who has the most collected painting in the MET? Picasso? Andy Warhol?
    - 1.3. The MET has experienced different directors since it was founded in 1870. Would the director drive the trend of collection type through the years?
    - 1.4. Private collection deep-dive
    - 1.5. If we take a deeper look into a collection by artist nationaly, for example, China, what are the collections? How are they displayed?

2. **With over 150 years of history, how has the Met evolved? Would it continue to hold the largest art museums in the history? We would like to take a closer look at where are those collections are from?**
    - 2.1. What is the source of collections (Gift / purchase)
    - 2.2. Who has been donating, and are donation concentrated?
    - 2.3. Most artwork are modern time art
    - 2.4. Accession quantity count by decade
    - 2.5. Comprehensive art collection for artist

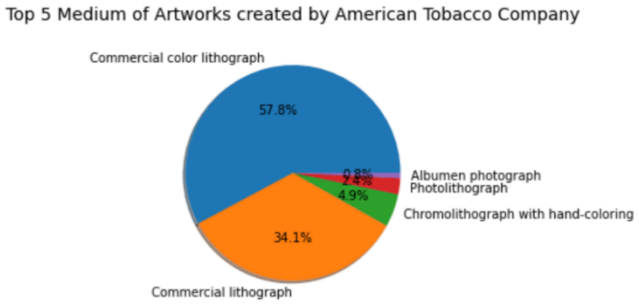
In this following section, we will be exploring our first question - **Is the Met the same with how it appears to be when we visit?** We will be assessing this question via breakdown hypothesis and questions, with answers, facts, data visualization, interpretation and analysis under each finding.

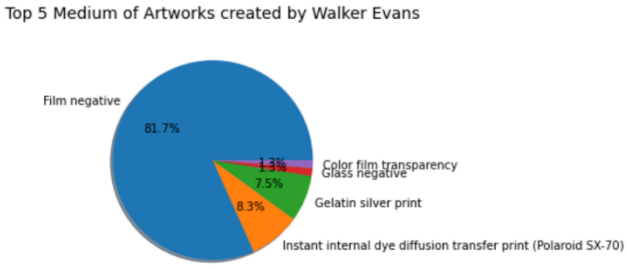
1.1 (Hypothesis) Top artists with most collections at the Met are always those commonly known ones.

From common sense perspective, we usually believe that top artists with most collections at MET are those names everyone is familiar with, however the visualization generated from Python below clearly indicates that this hypothesis is null. None of the top 10 artists are famous and many of them are even company names. To dive deeper, naturally two questions come up - 1) Is this result valid, why company names are on the top artist list? 2) Who is Walker Evans? Why he has so many collections at MET?



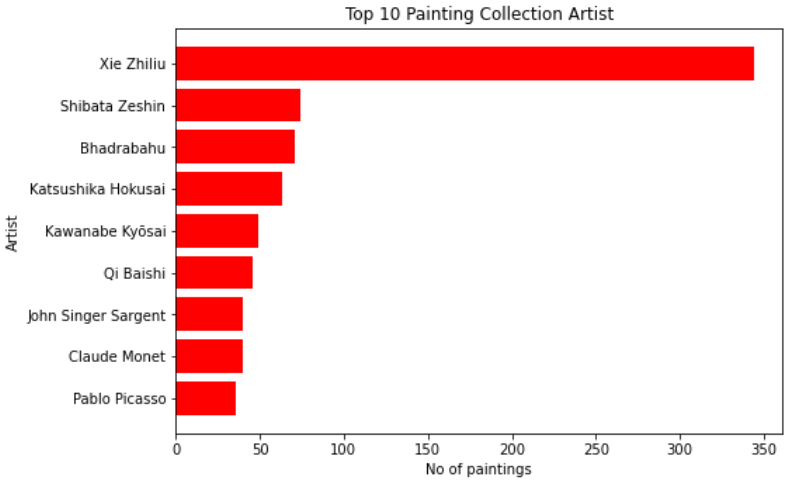
Through using the variable - **Medium** (type of art, such as photos, paintings, sculptures etc.), it effectively addresses the questions, majority of artwork by American Tobacco Company are commercial color lithograph. And most of Walker Evans artwork are Film negative (a type of photograph). As the creation of commercial graphs and photos are easier, the volume is high thus put them at the top artist list rather than those famous artists we know in the history.



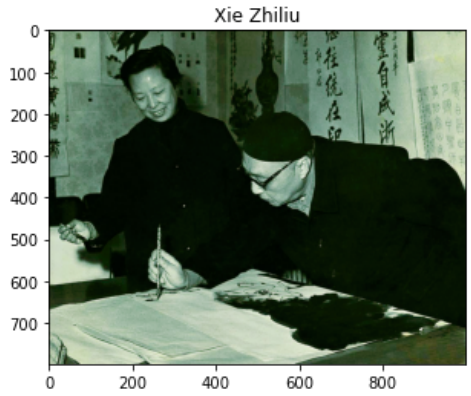


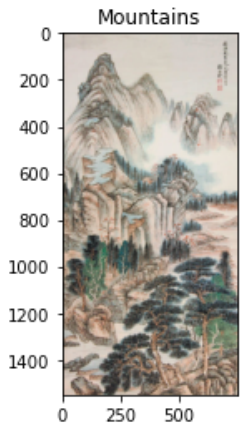
1.2 (Hypothesis) We all love to view paintings. Famous artists like Picasso and Andy Warhol should have the most painting collected in the MET

We created a subset data to include rows where ‘classification’ equal to ‘Paintings’. There are 10470 rows, accounts for 2.2% of the collection within the Met. By using groupby and value_counts, we are able to rank the top 10 artist name by their number of collections.



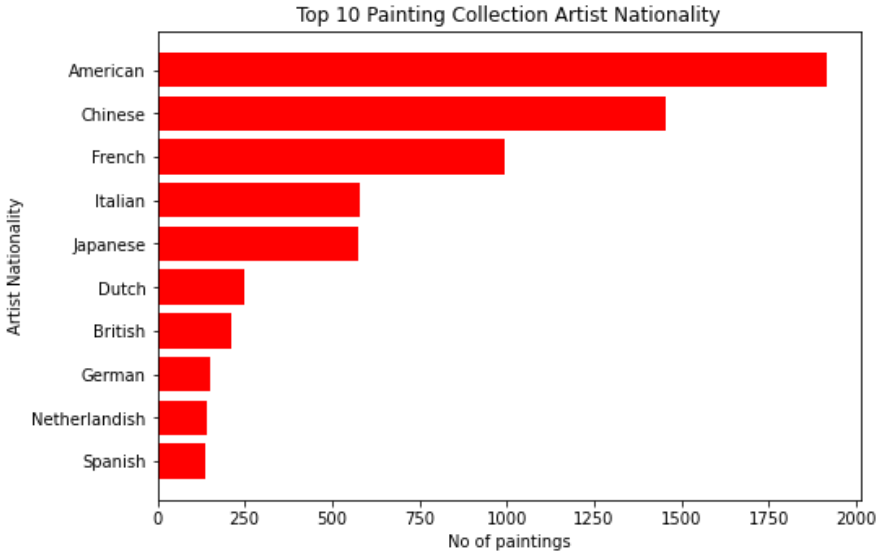
Surprisingly, the top artist by number of paintings are Mr. Xie Zhiliu from China (1917 – 1997) with 344 paintings. By researching further on this, the reason is Mr. Xie's daughter has donated over 150 paintings to the Met and the Met held a special exhibition featuring him in 2010. We use Python to download one of his paintings. Despite the surprising result of the artist of the most collected paintings, Monet and Picasso still make to the top 10 list.





In addition, what are the top artist nationality within paintings? Americans, Italian or French?

Similarly, we created subset dataset and use groupby to count by artist nationality. The result also surprised us a little bit. Indeed, American, Italian and French artist made to the top 5. The Met also has a huge collection of paintings from Asian countries including Chinese and Japanese.



1.3 (Hypothesis) The Met had 11 directors in its 151 years of history, and they have contributed significantly to its collection. We think those directors would exhibit different collection preferences through their respective periods.

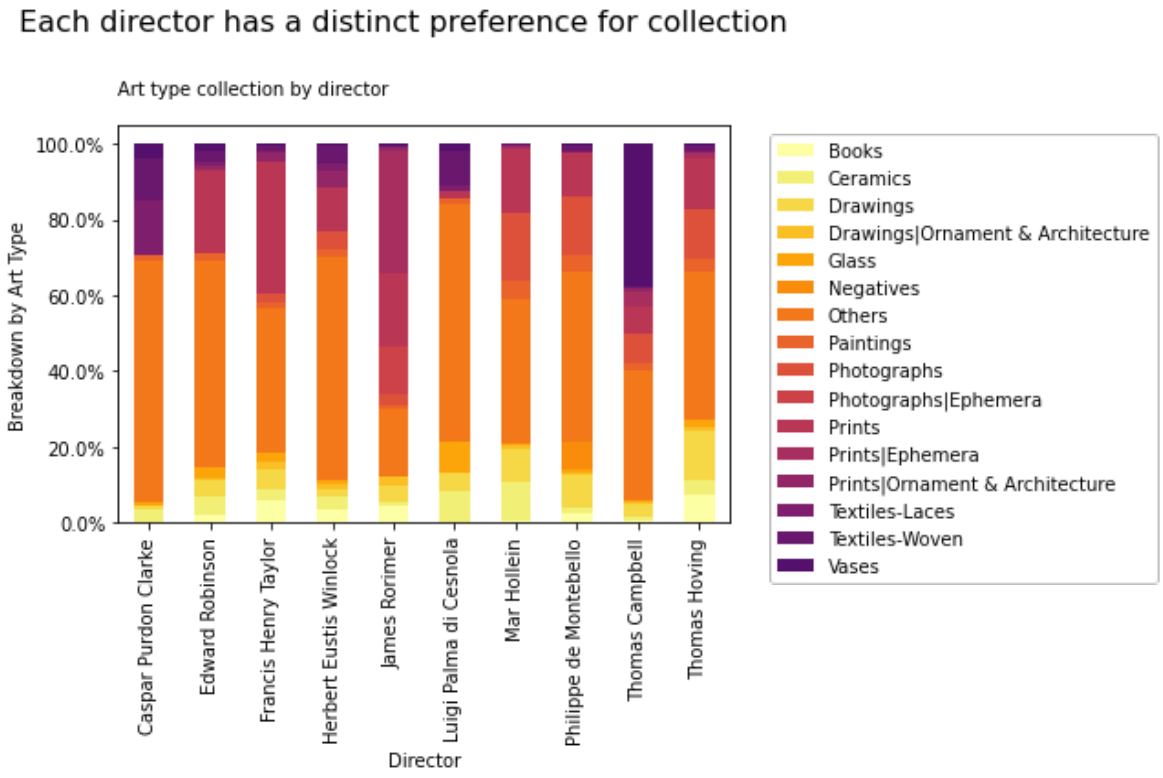
We use time-series data by art type by director period informed by the Met history.

1. Create a subset dataframe to include columns of object_id, accessionyear, classification
2. Select top 20 classification art types and rest as 'Others'. Create a new column called 'classification1' to store the modified art type
3. Create a new column called "director" to display the time series of accession year by director period

Below is the list of directors and their terms:

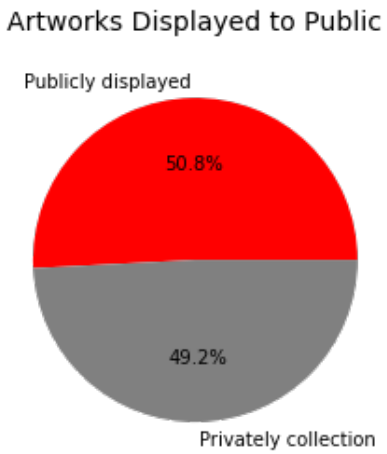
Names	Term
Luigi Palma di Cesnola	1879 to 1904
Caspar Purdon Clarke	1904 to 1910
Edward Robinson	1910 to 1931
Herbert Eustis Winlock	1932 to 1939
Francis Henry Taylor	1940 to 1955
James Rorimer	1955 to 1966
Thomas Hoving	1967 to 1977
Philippe de Montebello	1977 to 2008
Thomas P. Campbell	2009 to 2017
Daniel Weiss	2017 to 2018
Max Hollein	2018

The result shows each director has a distinctive preference for collection. We think this is aligned with our hypothesis and important for the diversification and richness of the collection.

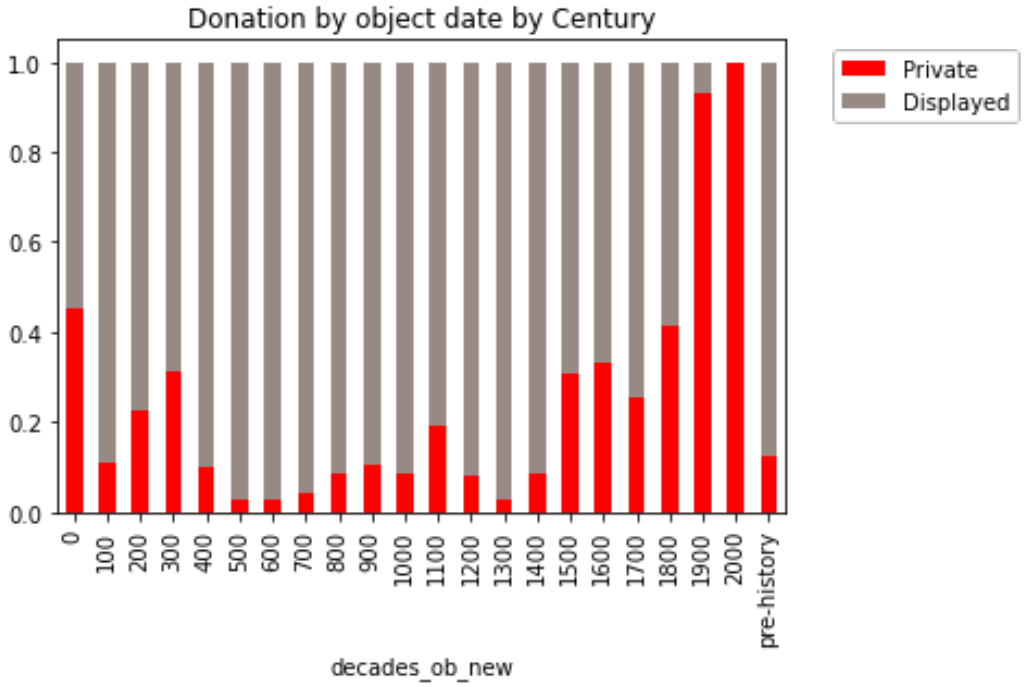


1.4 A deep-dive into private collections at the Met

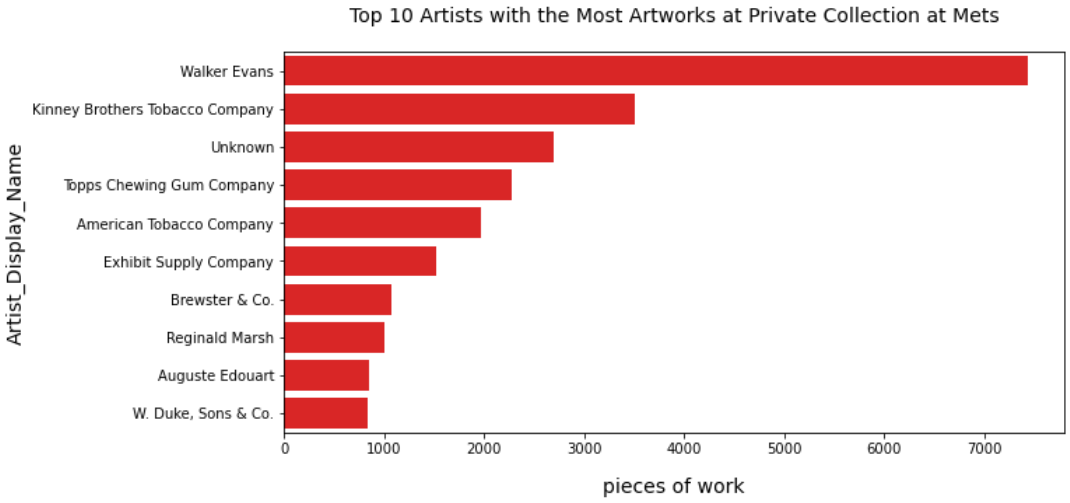
We have performed analysis to deep diving into the private collections, through using a combination of *groupby, value_counts, and other methods*. Among the entire collections, around half of the collections are displayed publicly.



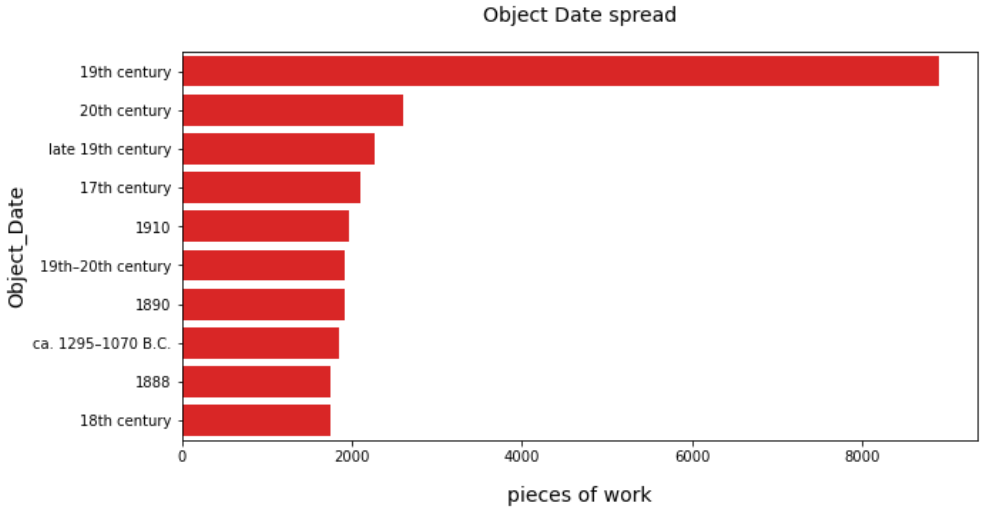
We have further breakdown the proportion of publicly displayed goods by vintage of the works. Seems that collections with vintage around A.D. 500 – 1300 are almost entirely displayed (~90%) to public. A fairly big percentage (>90%) of artworks dated in recent 2 centuries (1900 – 2000) are not displayed.



We have also ranked the top 10 artists whose work has been within private collection, through *sort_value function*. Our old friend Walker Evans tops the table again.

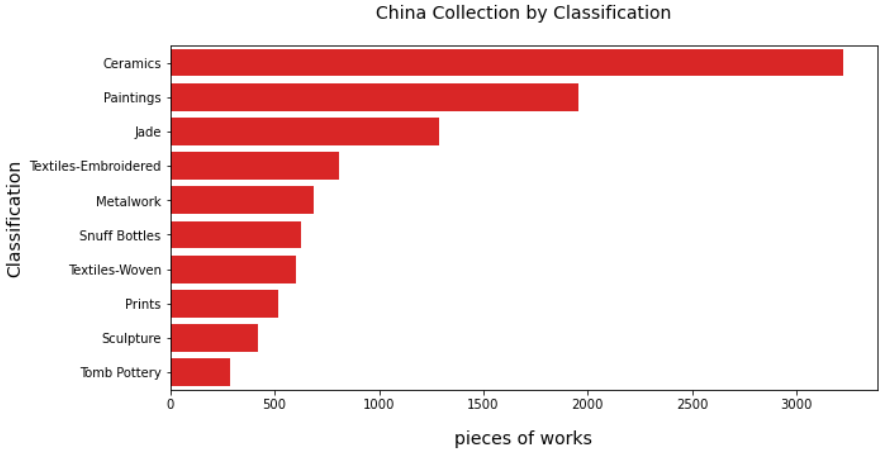


And works in 19th century composed most of the private collections.

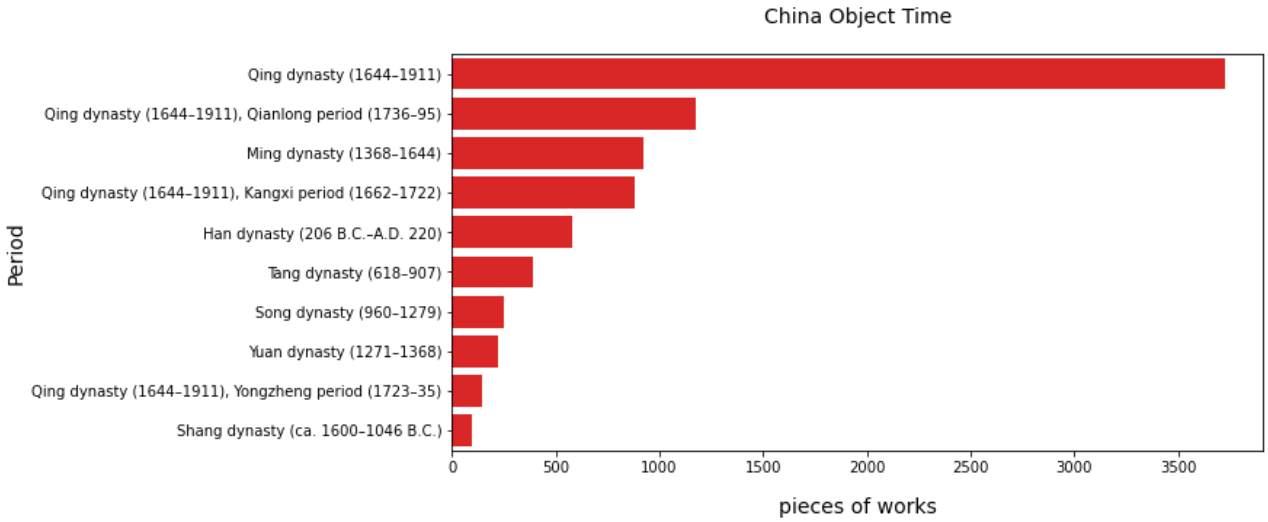


1.5 *If we take a closer look into a collection by artist nationaly, for example, China, what are the collections? How are they displayed?*

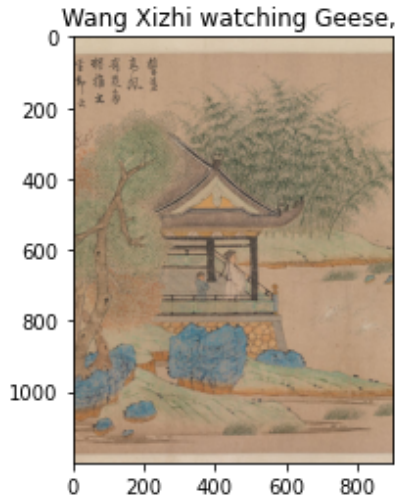
We have zoomed in a particular to collections categorized as Chinese culture . As we would expect, and if you are familiar with China history, most of the Chinese culture related collections are ceramics, which tops the collections within China collections (China also means ceramics), followed by paintings



Most of collections are from Qing Dynasty (most recent)

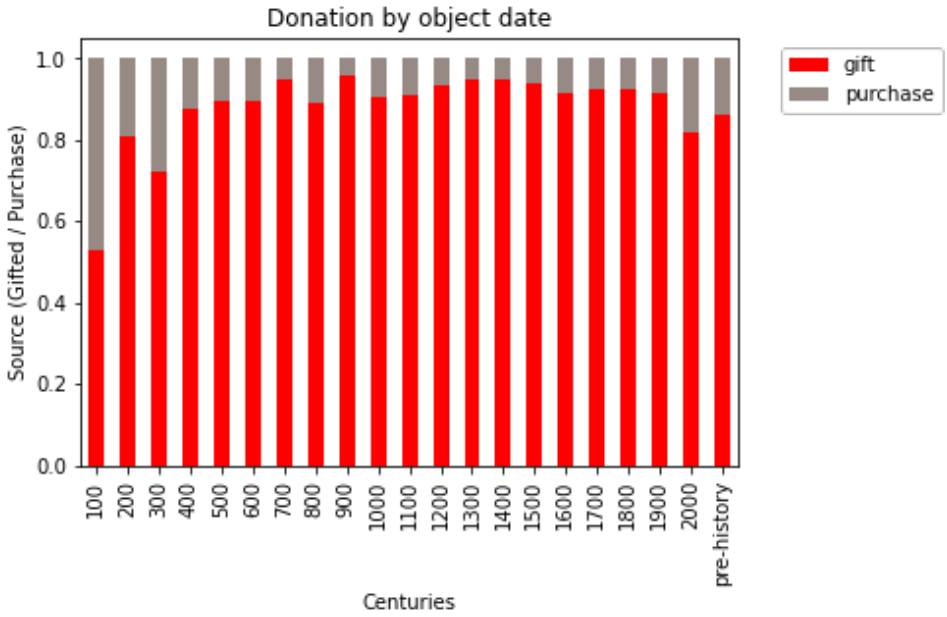


A typical famous painting at the Met: Wang Xizhi watching Geese (ca. 1295)

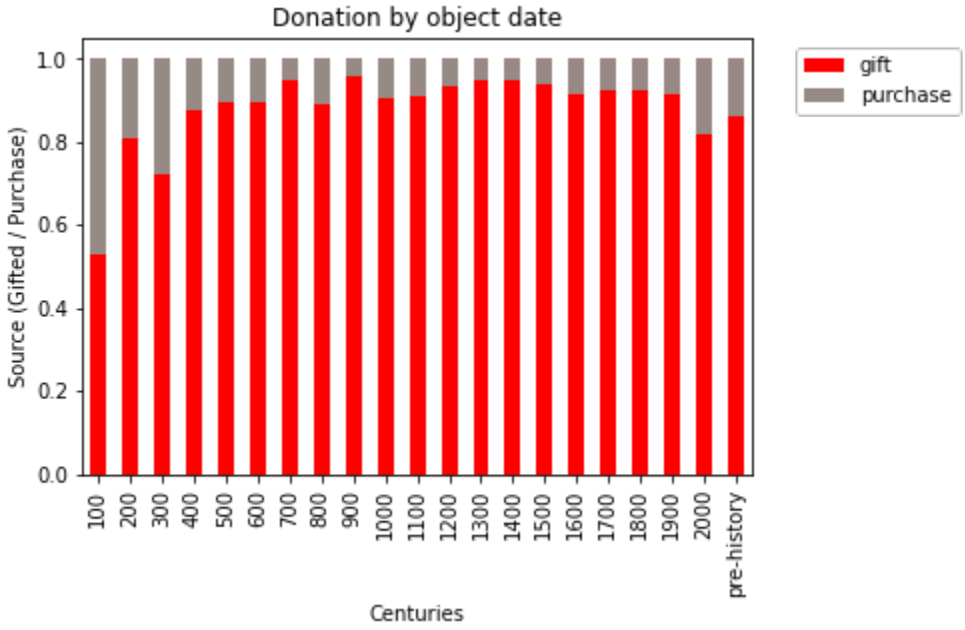


2.1 What is the source of collections (Gift / purchaseed)

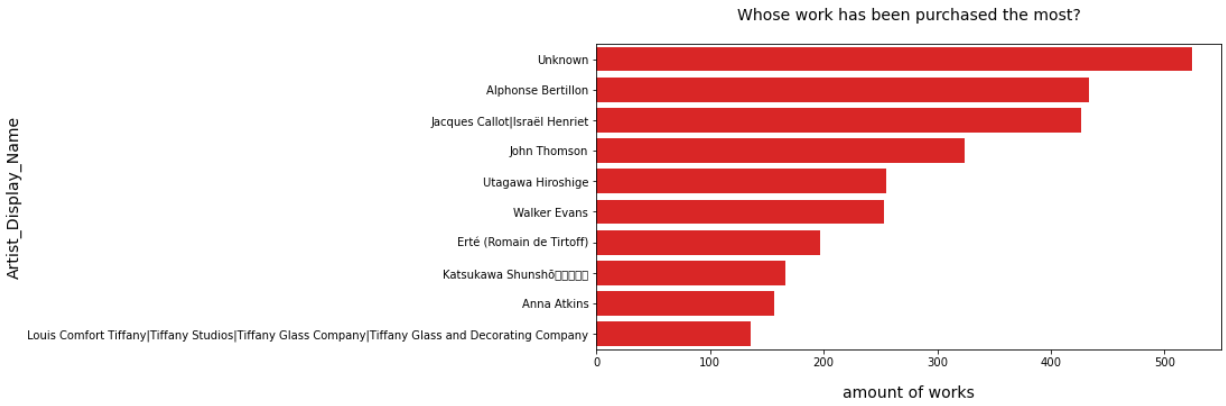
Objects at the Met are either gifted or purchase (sometimes through a swap). We have broken down the percentage of gifted and purchased goods, based on the centuries that objects are created. As evidenced in the charts below that the Met has started to purchase more contemporary goods.



In terms of purchase ratio (amount of collections being purchased during a period) more collections have been purchased than gifted recently. Anecdotally, 1870 is the period when Met is incorporated (which makes this data point an outlier). This could be a sign that, proportionally, fewer people is gifting to the Met. Which leads to another one of **our hypothesis that is people getting less enthusiastic into gifting to museum?**



Let's put aside the question whether people are getting less generous for a moment. An equally intriguing follow-up question is that whose work has the Met been purchasing? Below analysis and chart offers a glimpse into this question.



The artists whose work has been purchased the most by the Met is Alphonse Bertillon. Who is him?

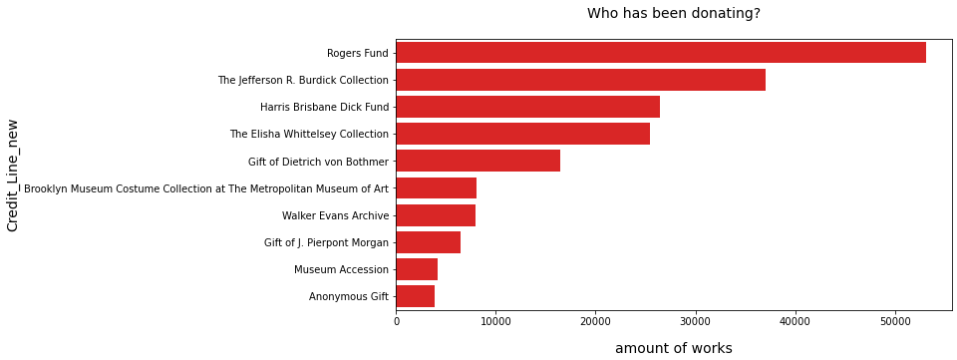
Turns out he is a French police officer and biometrics researcher who applied the anthropological technique of anthropometry to law enforcement creating an identification system based on physical measurements. Anthropometry was the first scientific system used by police to identify criminals. Before that time, criminals could only be identified by name or photograph. He is also the inventor of the mug shot [1]

Alphonse pioneers 's work tops the table by amount of works (or pictures of suspects he collected). Then we thought this analysis might be biased that number of works collected might not be truly representative in terms of the significance of works. Hence we have zoomed into paintings that the Met has purchased. Turns out that Shibata Zeshin, a Japanese painter in Edo period, is the artist whose works have been purchased mostly, through an analysis enabled by *sort_value* method.

Artist	Amount
Shibata Zeshin	60
Loren MacIver	9
James Peale	8
Simon Denis	7
Gyokuen Bonpō	1
Dai Wan	1
Hugh Bridport	1
Zhao Boju	1

2.2 Who has been donating, and are donation concentrated?

Data based analysis suggests that Rogers fund has been the most generous donor historically, in terms of pieces of object donated.



Roger's fund has been actively donating to the Met in early part of 19th century.

Year	Pc. Donate
1909	3614
1908	3077
1948	2524
1917	2432
1920	2304
1941	2143
1922	1804
1912	1623
1925	1567
1913	1522

s

As quick fact, Jacob Rogers was president of Rogers Locomotive and Machine Works and served as the company's president. The company eventually became the second most popular steam locomotive manufacturer in North America. Upon Rogers' death in 1901, he bequeathed the majority of his fortune, amounting to $8 million, to the Metropolitan Museum of Art in New York City. The Museum continues to acquire art works in his name through the "Jacob S. Rogers Fund." [2]

We have also run a similar analysis of gifted artworks in the most recent decades. Below are the top donors. As we can see the amount of pieces top donors gifts to the Met is fewer, compared with similar period in last century. **This further corroborated with our hypothesis that people seems to be less into donating.**

Gilman Collection	1605
Gift of Nanette B. Kelekian	887
Samuel Eilenberg Collection	738
Bequest of William S. Lieberman	710
Gift of Muriel Kallis Newman	472
Gift of Sarah Shay	428
Gift of the artist	421
Gift of Mr. and Mrs. Eugene V. Thaw	242
Herbert Mitchell Collection	234

Comprehensive art collection for artist

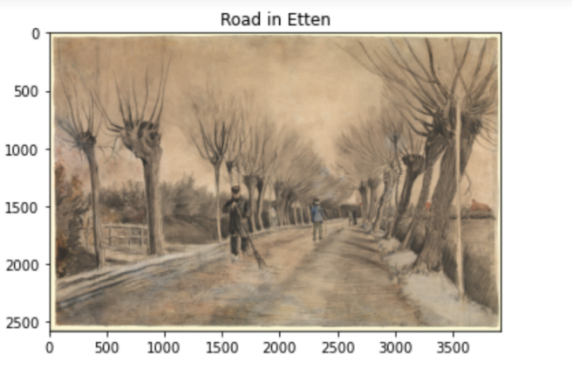
As the nature of our data comes along with the greatest visualization – Artwork, this is a bonus analysis by diving into two famous artists by pulling and organizing some their best work via Python.

Year 1888 is the breakthrough year of Van Gogh. *"The time in Arles became one of Van Gogh's more prolific periods: he completed 200 paintings and more than 100 drawings and watercolors.* ***He was enchanted by the local countryside and light; his works from this period are rich in yellow, ultramarine, and mauve. His paintings include harvests, wheat fields and general rural landmarks from the area."*** *- Wikipedia.*

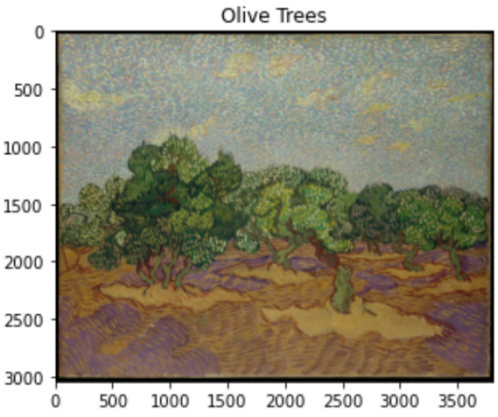
Question 1: Are we able to visually see this pre-1888 and post-1888 painting style change from Van Gogh's art?

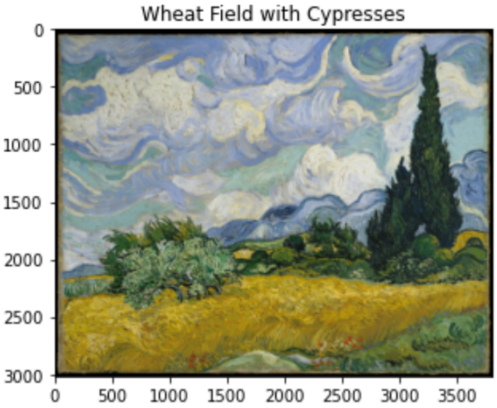
Step 1: Subset Van Gogh's artwork into two datasets - pre-1888 and post-1888
Step 2: Randomly select two from pre-1888 dataset and two from the post-1888 dataset.
Step 3: Use the URL from the dataset, to locate the artwork and pull in python. Using Pillow package to create mosaic style result.
Via randomly selected two pre-1888 pictures and post-1888 pictures, it can be clearly seen the style change, objects in those paintings are indeed harvest and rural landmarks as summarized by professionals.

Pre-1888



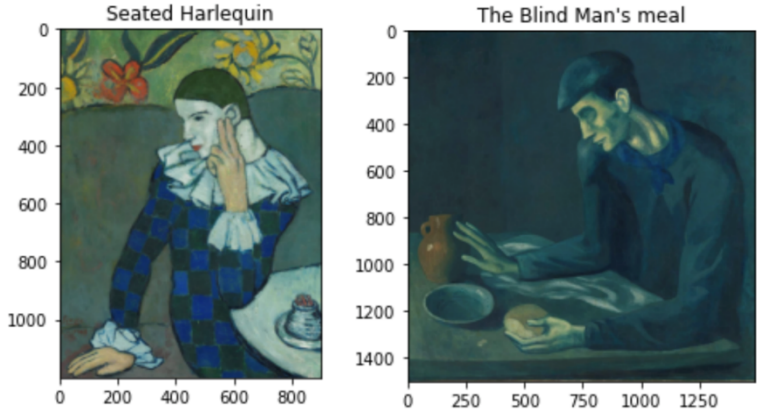
Post-1888





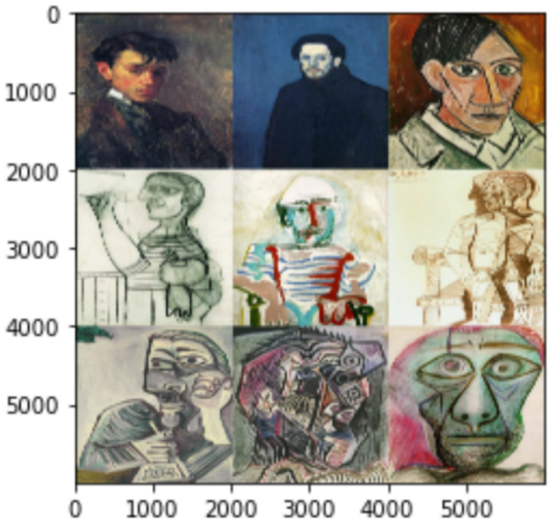
The second artist we would like to zoom in is Pablo Picasso. Picasso's work is often categorized into periods. One of them commonly accepted periods in his work are the Blue Period (1901–1904) - characterized by paintings rendered in shades of blue and blue-green only colors.

Here brings **Question 2: Is it true that his paining during 1901-1904 has a strong reflective of this Blue Period style where most colors should just be blue and blue-green?**
Similar as the processing steps listed above, we were able to randomly pick up few Picasso's artworks from this period and found the result exactly match with this hypothesis – indeed all of them are blue/blue-green color theme. See few examples below.



As well said by W. Somerset Maugham in the book The Moon and Sixpence, *"To my mind the most interesting thing in art is the personality of the artist",* here comes to our **3rd question - Can we find painting style change from Picasso's self-portrait?**

Assuming Picasso's constantly style changing is evident in his series of self-portraits, we located and organized his self-portrait painted from the age of 15 until 90, as vividly shown in a mosaic style pic below. We were fascinated by how clearly this one mosaic could clearly show his style change in each period, from realism, blue period, cubism till surrealism, each change is stunning and revolution.



Limitations

We hope the results would intrigue audience of the report to visit the Met again and select the collections to visit with a different perspective. In addition, it would also allow us to think further on the challenges of maintaining the status of such institution.

However, there are some limitations to our analysis. First of all, given the richness of the data, we could use others variables like department and artist gender to conduct more analysis but that is limited by time constraints.

In addition, some of the categorization displayed in the dataset is very technical. For example, under classification, there are different types and by various materials. That would make the categorization of the overall collection less straightforward to be understood by non-art savvy viewers.

Furthermore, the report uses a fairly amount of images to display the art work. Although the database shows the url link to the image from the Met, we can not open the image from the url due to restriction so we open the link with the image from internet for illustration.

Lastly, some of our analysis is based on amount of artworks; however, the dataset we have doesn't not offer any value of each artwork. Therefore, the result from the analysis might overstate and skew towards the quantity-based results, than value-based results.

Conclusion:

We think the Met has much more story to normal visitors like us than what it appears to be when we visit every time. We may go to Van Gaugh, Picasso and other famous artwork or artist immediately when we visit. However, the Met contains much more than that and it could be significantly impacted by where the piece of art come from.

In terms of whether the Met will be sustainable, our analysis shows two sides of arguments. On one hand, with the Med's collection heavily relies on gift and the gift donation has decreased in recent years. That might raise concern on whether the Met will enrich its collection continuously. On the other hand, a few other findings comfort us. Firstly, the Met keep acquiring diversified artist and art works through the years, which would help attract a wide range of visitors and contribute to the art development. In addition, the historical collection of the Met on important and popular artists are comprehensive (e.g. different self-portraits of Picasso). So it would keep attracting fans to visit and reinforce its status as one of the most important art museums in US or in the world.

Source

1. https://www.metmuseum.org/toah/hd/evan/hd_evan.htm
2. https://en.wikipedia.org/wiki/Vincent_van_Gogh