Project 2 Proposal: 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.

Key topics

We hope our analysis would cover below key topics and address below key questions. We intend to (1) present some general stats of the collections to give audience an overview of the collections, and (2) present our detailed findings on a few sub-category and specific artist's work at the Met through a deep-dive on related data

- General topics
  - Most popular artists
  - How is the vintage distribution of most of the collections?
  - What is in the "private collections"? Who contributed to them and their vintage thereof?
  - How has the collection trend evolved? And what are the drivers?
    - Time series of collection type
    - Time series of collection by region
  - In terms of source, how is the portion of gift vs. purchase changing over time
- Deep-dive
  - Painting and sculpture section
    - Who are the most collected artists?
  - Chinese collection section
    - by period, by artist, by type and by source
  - Vincent van Gogh
    - Which time period is his works collected the most by the Met
    - What are the specifics of his work at Met

Dataset description

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>

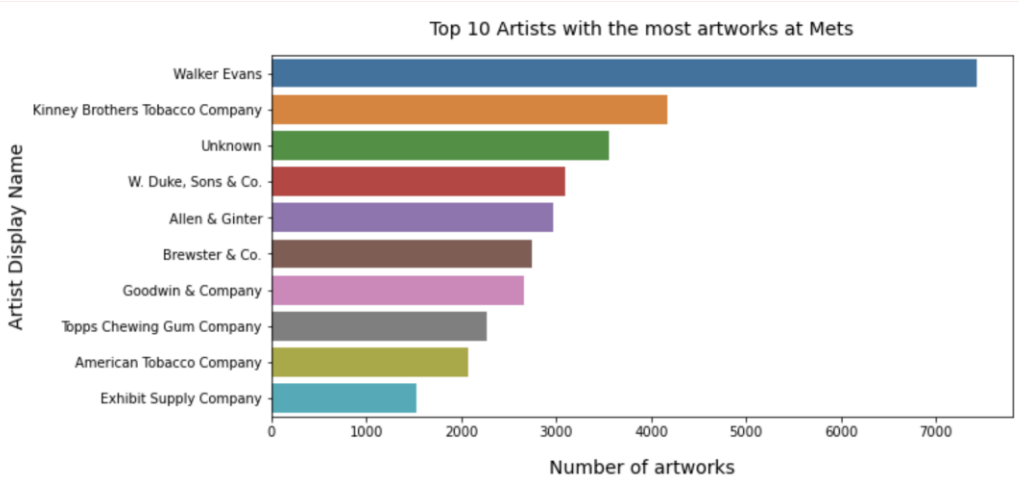
The csv file has 475,125 rows and >50 columns. Please refer to below "Key variables & insights" section for details.

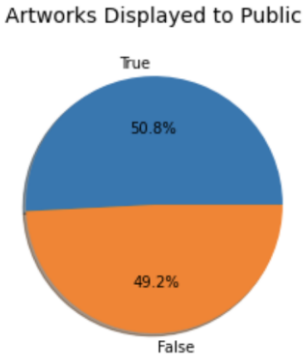
Approach

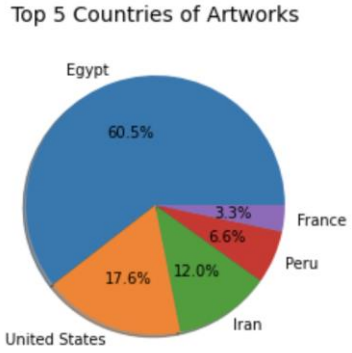
First, we will find out columns with missing values, to ensure the variables required for our analysis are in a

relatively high quality in terms of completeness and cleanness. Secondly, create the framework, generate visualization from selected variables per our key questions. And finally adding interpretations.

Initial plots, figures, or tables







Key variables & insights

There are 53 variables in this dataset, 10 variables have no missing values and 21 variables have less than 50% missing values. 20 key variables will be used for the analysis including but not limited to *Object Number, Country, Period, Dynasty, Reign, Artist Display Name, Artist Nationality, Object Date, is Public Domain, Object Name, Object Begin Date and Object End Date etc.*

Insights to be gained can be categorized into two types as below, which can hardly be recognized when visit the Mets and look at collections at the granular level.

- An overview of art collections by country/region, artist, public/private, gift/purchase
- An overview of time series of collection by region/type

Final deliverables

Our final report will be starting with 1) an overall project logistics, motivations/questions and goals; 2) A brief description of what key Python techniques used, and challenges faced; 3) The main content of the report will be data visualizations with detailed interpretation and analysis.