Artwork Master Database

The following document describes the contents of the Artwork Database and the process through which the Database was compiled.

The data was collected for the research of Renée B. Adams at the Saïd Business School, University of Oxford. Professor Adams’ research focuses on gender and its’ effects on the pricing of art on the primary markets.

<https://www.renee-adams.com>

<https://www.sbs.ox.ac.uk/about-us/people/renee-b-adams>

Table of contents

1. General description of the database
2. Process of data collection
3. Variable descriptions
4. Codes

1) General description of the database

The database contains artwork price data collected from the primary art market.

Due to Covid-19 most fairs during 2020 and 2021 were held online, which made the digital data collection possible. The majority of artworks in the database originate from [artsy.net](https://www.artsy.net), who have been hosting online art fairs since 2013. We have been actively collecting data since July 2020, however we also obtained data from fairs that took place earlier, thanks to artsy maintaining an archive.

We compiled a large database of artists and their gender from various sources, the primary being [artfacts.net](https://artfacts.net/). They provide a large range of data and detailed descriptions of artists and their works. Our primary interest was identifying the gender of the individual, however Artfacts provided us with a much wider range of variables, that can be utilised in the models.

Most of the data collection was automated via Python and falls under the U.K. Government’s “Exemptions to copyright” guideline as non-commercial research. [(for further reference)](https://www.gov.uk/guidance/exceptions-to-copyright%23content)

The current database contains 484,296 observations of which 189,550 (39.14%) are unique artworks and 179,052 (36.97%) are unique artworks from fairs. The duplicates are either due to us recording some fairs on multiple occasions or some artworks appearing at different fairs. We use the URL of an artwork to identify it as unique. We also assigned each observation a unique ID with which we can link it to the image of the artwork.

2) Process of data collection

**Art Basel 2020 was the first and only fair where we used manual downloads to extract price data from online art fairs.**

**Art Basel 2020** provides online viewing rooms for each exhibitor (https://www.artbasel.com/viewing-rooms). Upon entering each viewing room, you can browse displayed artwork one by one, and you can expand the webpage to see detailed information about the artwork.

The steps for collecting data:

1. Enter viewing rooms, open the webpage for each artwork, and download the webpages in the HTML format. The entire webpages were saved for the first time on the 2nd and the 3rd days of the art fair, and again on the 6th and 7th days. 3947 webpages (i.e. artworks) were saved at the first time and 3924 at the second.
2. Almost all the detailed information about an artwork is contained in the HTML file, from which I programmed in Python to extract. An essential program module I used is re, which provides regular expression matching operations. The only information not recorded in the HTML file is sector, and I added this information afterwards by matching galleries with sectors.
3. After running Python to generate the csv files, add the csv files into Excel using UTF-8 encoding.

**After Art Basel we decided to automate the webpage downloading. This way much manual labour can be saved and frequent saves can be made without much effort.**

**Masterpiece 2020** provided access to their viewing room through Artsy (https://www.artsy.net/masterpiece-online-2020). The website is only partially loaded when opened and artworks are grouped by galleries and are listed under each other. If the gallery has more than 15 artworks on display, there is a “See {} more artworks” button which has to be clicked to reveal the additional artworks. This fair could be downloaded automatically by using a Python code.

The program’s steps for collecting data:

1. Scroll to the bottom of the page to pre-load each gallery.
2. Scroll back to the top.
3. Repeat for each gallery:
4. Check whether the “See {} more artworks” button exists. This checks whether all the artworks are revealed.
5. If there is no button, proceed to the next gallery. Else click the button until all the artworks are revealed. (using a while loop: while the button exists -> click)
6. Save the HTML code of the fair and export the artwork URLs.
7. Download the HTML files corresponding to the URLs. The entire webpages were saved each other day from June 27 to July 9 (the fair was held between Jun 22nd – Jul 8th). Around 3240 ~ 3250 webpages (i.e. artworks) were saved each time.
8. Export the price data and other information of each artwork from the downloaded HTML files using the Python code.

**Other fairs on Artsy:** There are 416 fairs from 2013 to end of 2020 on Artsy. The same methodology as used for Masterpiece 2020 is applied to get price data for all those past Artsy fairs.

**Contact information of galleries**: For galleries that attended ArtBasel 2020 or any of the Artsy art fairs until Aug 2020, their contact information is collected for the purpose of survey.

For the galleries whose website were in Chinese the contact information had to be collected manually. This meant going to their websites’ and searching for an email address and phone number. If the information could not be found on the website, other sources (like Facebook, LinkedIn, etc.) had to be located and searched.

**Vienna Contemporary 2020:** The fair opened on the 22nd of September. We made saves daily starting September 23rd up until days after the fair. In the first days there were only approximately 250 artworks, however as the fair progressed, we got more substantial data. There were no issues with writing the code that performed the automated data extraction.

The automated process of the data extraction:

1. We login to the viewing room and accept cookies.
2. All artworks are featured on the main page. The program only has to click on the ‘Show More’ button to expand the viewing room. Once the button ceases to exist the page can be considered fully loaded.
3. We identify each artwork using their unique XPATH and save the URL of the artworks. We use a loop to identify each artwork based on their unique integer ID.
4. Open each page and extract the data displayed.

**Frieze Masters 2020 & Frieze London 2020:** These two fairs were hosted in the same viewing room. We made the first save on the 10th of October and kept making daily saves until the end of the fair on the 15th of October. Due to the layout of the webpage and the lack of structure within the data there were issues separating parameters displayed on the website. This must be done after the save is complete using the regex package in python. Additionally, we also collected gallery contact data from the fair. We used this dataset to expand our gallery masterlist.

The automated process of the data extraction:

1. We identified that each gallery has a unique URL, with a unique integer at the end. (https://viewingroom.frieze.com/viewing-room/ID).
2. Open each gallery based on this URL, looping through all possible IDs, until we reach an ID that does not correspond to a gallery.
3. From the gallery page extract the gallery contact information.
4. We loop through their artworks that are displayed on the bottom of the page and collect their URLs.
5. Open each artwork and extract the data displayed.

**Art Basel-OVR 2020:** This fair ran over the weekend of the 25th and 26th of September. Due to the short timeframe we had to pay attention that code was working properly. Since the format of the website differed from those that we are mainly used to, we had to make major modifications to the code. We used a similar approach at this fair as before. The data was collected automatically. We managed to extract accurate price data for all 600 artworks displayed at the fair.

The automated process of the data extraction:

1. Sort artworks based on artist so that we can validate results in the extracted data.
2. Scroll to the bottom of the webpage to pre-load the viewing room.
3. Identify artworks based on their unique XPATH and save their URLs.
4. Open the website of each artwork and save the data displayed.

**Artsy Fairs:** While we are keeping track of major art fairs, there are also some smaller fairs featured on Artsy. We also save price information from these to maximise the size of the dataset and reduce the selection bias. The number of artworks at these fairs vary. We also do not make daily saves as we do not anticipate as much movement/change in prices as at major fairs.

These fairs include 1-54 London, Apertura Madrid Gallery Weekend 2020, Barcelona Gallery Weekend 2020, Intersect Aspen, Latitude 2020, Positions Berlin 2020, etc.

The code has gone through many changes over time. Before IPFDA 2020 (before 22/12/2020) the code used the following logic to extract data:

1. For each gallery identify and click on the ‘See {x} More Artworks’ button (if available).
2. Save the URLs of every artwork.
3. Download each artwork’s website’s HTML code using multiprocessing and extract the data displayed.

However, when we attempted to use this code on IPFDA 2020 the Artsy website would crash. This forced us to make the following changes.

1. Keep track of the number of galleries whose viewing rooms have been expanded.
2. For a range of galleries in the total, identify and click on the ‘See {x} More Artworks’ button (if available).
3. Save the URLs of every artwork on display.
4. Reload the page and go back to Step 2, until all the galleries’ artworks are seen.
5. Download each artwork’s website’s HTML code using multiprocessing and extract the data displayed.

**Price Requests:** In order to obtain price information for those artworks whose prices were not publicly displayed we decided to reach out to galleries. There were two main routes that we considered taking:

1. Making requests through Artsy (specifically for those artworks that were displayed there)
2. Sending out survey emails to galleries

(1) Artsy Requests:

With this process we faced many issues. The first code that we ran on the website requested price information to circa 10,000 artworks and got only 2 responses. After this failure we realised that a real person would not be able to send 10,000 requests in a matter of hours and that the website might log the user as a bot and nullify its requests.   
We decided to conduct a test to see whether the lack of responses were indeed due to this error.   
No response from galleries after inquiries. Either they do not use Artsy’s inquiry system or the messages get flagged.

3) Variable descriptions

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type** | **Description** |
| artwork\_url\_id | Float | *group(url)*  Consistent throughout fairs and saves. |
| image\_download\_save\_id | Float | ID assigned when downloading images. No consistency throughout fairs or saves |
| image\_id | String | ID to match artwork 1:1 to the downloaded image based on its filename. |
| is\_fair | Dummy/Byte | Marks whether the observation was collected from an art fair.  0: non-fair data 1: fair data |
| artsy\_data | Dummy/Byte | Marks whether the observation was collected from artsy.  0: non-artsy data 1: artsy data |
| artfacts\_data\_available | Dummy/Byte | Marks whether the observation’s artist could be matched to the artfacts artist database.  0: no match 1: match |
| title | String | The title of the artwork. |
| year\_completed | Integer | The year when the artwork was completed. |
| years | Integer | The number of years taken to create the artwork. |
| bc\_dummy | Dummy/Byte | Marks whether the years are BC.  0: AD 1: BC |
| price | Float | *(low\_price\_USD + high\_price\_USD) / 2*  The mean of low & high end range converted to USD. |
| ln\_price | Float | *log(price)*  The natural log of price. Useful for analysis since prices are log-normally distributed. |
| low\_price\_USD | Float | *low\_price \* exchange\_rate*  The lower end of the price range converted to USD. |
| high\_price\_USD | Float | *high\_price \* exchange\_rate*  The higher end of the price range converted to USD. |
| low\_price | Float | The lower end of the price range specified in the original currency. |
| high\_price | Float | The higher end of the price range specified in the original currency. |

|  |  |  |
| --- | --- | --- |
| currency | String | The 3 letter abbreviation of the currency in which the artwork was originally advertised. |
| category | Integer | Categorical variable. |
| category\_inferred | Integer | The same variable as category, with additional observations’ classification inferred based on common materials in each group. |
| materials | String | The materials used in the creation of the artwork. |
| length\_in | Float | The length of the artwork in inches. |
| width\_in | Float | The width of the artwork in inches. |
| depth\_in | Float | The depth of the artwork in inches. |
| sold | Dummy/Byte | Marks whether the artwork was sold or not.  0: not sold 1:sold |
| sold\_inferred | Dummy/Byte | The same variable as sold, with additional observations’ classification inferred based on the sold status of other artworks within the fair. |
| collection\_date | Date (Integer) | The date the artwork’s data was collected in Stata date format. |
| collection\_times | Integer | The number of times an artwork appears in the database. |
| collection\_date1 | Date (Integer) |  |
| collection\_date2 | Date (Integer) |  |
| collection\_date1\_fair | Integer |  |
| collection\_date2\_fair | Integer |  |
| sold\_date | Date (Integer) | The date the artwork was sold. |
| exit\_date | Integer |  |
| gallery\_id | Integer | *group(gallery)*  The unique identifier of a gallery based on its name. |
| gallery | String | The name of the gallery representing the artist. |
| gallery\_slug | String | The artsy slug of the gallery. |
| fair\_id | String | The ID of fair. |
| fair\_name | String | The name of the fair. |
| fair\_image\_id | String | A combination of the fair\_id and fair\_year variables such that fairs in different years may be distinguished for the purpose of downloading images. |
| fair\_year | Integer | The year of the fair. |
| number\_of\_artworks | Long | The number of artworks that appeared on the fair. |
| fair\_start\_date | Date (Integer) | The date the fair started in Stata format. |
| fair\_end\_date | Date (Integer) | The date the fair ended in Stata format (if applicable). |
| artist\_id | Integer | *group(artist)*  Unique identifier of an artist based on their name. Consistent throughout fairs. |
| artist | String | The name of the artist |
| birth\_year | Date (Integer) | The birth year of the artist based on various sources. |
| death\_year | Date (Integer) | The death date of the artist based on various sources (if applicable). |
| artist\_slug |  |  |
| acclaim |  |  |
| gender |  |  |
| gender\_source |  |  |
| artist\_g7 |  |  |
| artist\_advanced\_econ |  |  |
| artist\_eu |  |  |
| artist\_first\_nationality |  |  |
| artist\_secondary\_nationality |  |  |
| artist\_dual\_nationality |  |  |
| inconsistency |  |  |
| manual\_correction |  |  |
| artist\_data\_available |  |  |
| aF\_gender |  |  |
| au\_gender\_MANUAL |  |  |
| au\_gender |  |  |
| bl\_gender |  |  |
| ar\_gender |  |  |
| nn\_gender |  |  |
| exact\_match\_gender |  |  |
| aF\_birth\_date |  |  |
| aF\_birth\_year |  |  |
| aF\_death\_date |  |  |
| aF\_death\_year |  |  |
| aF\_birth\_location |  |  |
| aF\_nationality |  |  |
| aF\_countries |  |  |
| aF\_sector |  |  |
| aF\_ranking |  |  |
| aF\_exhibitions\_in\_country\_1 |  |  |
| aF\_exhibitions\_in\_count\_1 |  |  |
| aF\_exhibitions\_in\_country\_2 |  |  |
| aF\_exhibitions\_in\_count\_2 |  |  |
| aF\_exhibitions\_in\_country\_3 |  |  |
| aF\_exhibitions\_in\_count\_3 |  |  |
| aF\_exhibitions\_at\_institution\_1 |  |  |
| aF\_institution\_id\_1 |  |  |
| aF\_exhibitions\_at\_institution\_2 |  |  |
| aF\_institution\_id\_2 |  |  |
| aF\_exhibitions\_at\_institution\_3 |  |  |
| aF\_institution\_id\_3 |  |  |
| aF\_national\_ranking |  |  |
| aF\_exhibitions\_total |  |  |
| au\_derived\_nationality |  |  |
| au\_artist\_birth\_place |  |  |
| au\_artist\_year\_birth |  |  |
| au\_deceased |  |  |
| bl\_nationality |  |  |
| bl\_yobirth |  |  |
| bl\_yodeath |  |  |
| birth\_error |  |  |
| death\_error |  |  |
| price\_unformatted |  |  |
| fair |  |  |
| filename |  |  |
| year |  |  |
| start\_year |  |  |
| end\_year |  |  |
| century\_dummy |  |  |
| cent\_early |  |  |
| cent\_late |  |  |
| ln\_low\_price\_USD |  |  |
| ln\_high\_price\_USD |  |  |
| datecur |  |  |
| exchange\_rate |  |  |
| dimensions |  |  |
| length |  |  |
| width |  |  |
| depth |  |  |
| measurement |  |  |
| length\_cm |  |  |
| width\_cm |  |  |
| depth\_cm |  |  |
| url |  |  |
| image\_url |  |  |
| collection\_url |  |  |
| sold\_unformatted |  |  |
| collectiondate\_unformatted |  |  |
| details |  |  |
| description |  |  |
| \_dupe |  |  |
| fair\_date |  |  |
| fair\_artsy\_url |  |  |
| aF\_id |  |  |
| aF\_first\_name |  |  |
| aF\_last\_name |  |  |
| aF\_aliases |  |  |
| aF\_birth\_date\_unformatted |  |  |
| aF\_death\_date\_unformatted |  |  |
| aF\_death\_location |  |  |
| aF\_movements |  |  |
| aF\_media |  |  |
| aF\_link\_artfacts |  |  |
| aF\_link\_personal |  |  |
| aF\_links\_wiki |  |  |
| aF\_ranking\_trend |  |  |
| aF\_cities |  |  |
| aF\_artists |  |  |
| aF\_groups |  |  |
| aF\_institution\_1\_link\_artfacts |  |  |
| aF\_institution\_1\_link |  |  |
| aF\_institution\_1\_link\_facebook |  |  |
| aF\_institution\_1\_link\_twitter |  |  |
| aF\_institution\_1\_link\_wiki |  |  |
| aF\_institution\_2\_link\_artfacts |  |  |
| aF\_institution\_2\_link |  |  |
| aF\_institution\_2\_link\_facebook |  |  |
| aF\_institution\_2\_link\_twitter |  |  |
| aF\_institution\_2\_link\_wiki |  |  |
| aF\_institution\_3\_link\_artfacts |  |  |
| aF\_institution\_3\_link |  |  |
| aF\_institution\_3\_link\_facebook |  |  |
| aF\_institution\_3\_link\_twitter |  |  |
| aF\_institution\_3\_link\_wiki |  |  |
| aF\_description |  |  |
| aF\_exhibitions\_solo |  |  |
| aF\_exhibitions\_group |  |  |
| aF\_exhibitions\_artfair |  |  |
| aF\_exhibitions\_collective |  |  |
| aF\_exhibitions\_current |  |  |
| aF\_exhibitions\_biennial |  |  |
| exhibitions\_region\_eastafrica |  |  |
| exhibitions\_region\_middleafrica |  |  |
| exhibitions\_region\_northafrica |  |  |
| exhibitions\_region\_southafrica |  |  |
| exhibitions\_region\_westafrica |  |  |
| exhibitions\_region\_caribbean |  |  |
| exhibitions\_region\_centralameric |  |  |
| exhibitions\_region\_southamerica |  |  |
| exhibitions\_region\_northamerica |  |  |
| exhibitions\_region\_centralasia |  |  |
| exhibitions\_region\_eastasia |  |  |
| exhibitions\_region\_southasia |  |  |
| exhibitions\_region\_southeastasia |  |  |
| exhibitions\_region\_westasia |  |  |
| exhibitions\_region\_easteurope |  |  |
| exhibitions\_region\_northeurope |  |  |
| exhibitions\_region\_southeurope |  |  |
| exhibitions\_region\_westeurope |  |  |
| exhibitions\_region\_australianewz |  |  |
| exhibitions\_region\_melanesia |  |  |
| exhibitions\_region\_polynesia |  |  |
| exhibitions\_region\_micronesia |  |  |
| aF\_exhebitions\_museum |  |  |
| aF\_exhebitions\_gallery |  |  |
| aF\_exhebitions\_artfair |  |  |
| aF\_exhebitions\_biennial |  |  |
| aF\_exhebitions\_other |  |  |
| aF\_catalog |  |  |
| aF\_dealer |  |  |
| aF\_collection |  |  |
| ar\_url |  |  |
| ar\_artsy\_prediction\_dummy |  |  |
| ar\_description |  |  |
| ar\_artsy\_female\_counter |  |  |
| ar\_artsy\_male\_counter |  |  |
| ar\_artsy\_prediction |  |  |
| au\_artist\_surname |  |  |
| au\_artist\_forename |  |  |
| au\_example\_artwork\_title |  |  |
| au\_medium |  |  |
| au\_nationality\_code |  |  |
| au\_artist\_id |  |  |
| au\_artwork\_id |  |  |
| au\_logsurface |  |  |
| au\_marked |  |  |
| au\_style |  |  |
| bl\_malepronouns |  |  |
| bl\_femalepronouns |  |  |
| bl\_predictionaccuracy |  |  |
| bl\_url |  |  |
| exact\_match |  |  |
| nn\_valid\_input |  |  |
| nn\_value |  |  |
| nn\_accuracy |  |  |
| sector |  |  |
| category\_missing |  |  |
| fair\_group |  |  |