**DATA**

I use three datasets on auction sales in this paper: Impressionist art (1980-1991), Contemporary art (1982-1994), and recent assorted art sales (2006-2016). The Impressionist and Contemporary art datasets have been used extensively in the literature[[1]](#footnote-1) and are already described in detail elsewhere[[2]](#footnote-2). Both of those datasets are available on the Brandeis website of Kathryn Graddy[[3]](#footnote-3). However, the last is a new dataset constructed specifically for this paper.

**IMPRESSIONIST ART (1980-1991)**

The Impressionist art dataset (1980-1991) was constructed by Orley Ashenfelter and Andrew Richardson in 1992, and covers sales at Christie’s and Sotheby’s in both London and New York. There are well over 16,000 observations of art piece sales, which were compiled by manually scouring presale catalogs. Each observation contains the painting title, the artist name, the sale price and date, the auction house and location, the presale low and high estimates, and hedonic characteristics such as the piece dimensions and the presence of a signature. The dataset contains 58 major artists whose work is often featured at auction, and among the most frequent are Pablo Picasso (1881-1973), Raoul Dufy (1877-1953), and Pierre Renoir (1841-1919). Approximately half the auction sales are split between Christie’s and Sotheby’s, as well as between London and New York. Table 1 shows summary statistics for selected attributes. The highest sale in this dataset goes to Vincent Van Gogh’s *Portrait of Dr. Gachet* (1890), which netted approximately $82.5 million on May 15, 1990. Conversely, the lowest sale is a work by Paul Cesar Helleu that went for a mere $1,888 on March 25, 1986.

Many of these quantities have distributions that are roughly log-normal (i.e., without the log transformation, skewed heavily right), showed in Figure 1. In general, we see very high variation: sales price, for example, reflects both paintings with record-high sales, as well as paintings that sold for minimal amounts or were bought in. This is because the majority of paintings exhibit middle-market sale price, estimates, size, and so forth, while relatively few reach the highest ranges. The two painting dimensions have the most irregular distributions, particularly in the middle ranges. However, as seen in Figure 2, the large portion of paintings do not tend to be lopsided with respect to their physical dimensions. Finally, we see that auction sales in this dataset have tended to grow over time (Figure 3), though there are clearly some huge years with record numbers of sales.

**CONTEMPORARY ART (1982-1994)**

Next, the Contemporary art dataset represents every Contemporary art piece sold from 1982 to 1994 at Christie’s primary King Street location in London, for a total of approximately 4,500 observations. The dataset was compiled by Kathryn Graddy, who manually examined auction catalogs and sifted through internal data in the archives of Christie’s. Similar to the Impressionist dataset, each observation lists the artist, the auction sale price and date, the presale low and high estimates, the lot number, whether or not the item sold, and hedonic characteristics such as the artist and medium. Various currency exchange quantities are included, such as the UK CPI at the time, and monetary quantities are given in thousands of pounds. Nearly 600 artists are represented, with Lucio Fontana (1899-1968), Karel Appel (1921-2006), and Alexander Calder (1898-1976) being the most frequent.

Table 2 gives summary statistics for this dataset. As with the Impressionist dataset, and shown in Figure 4, quantitative dimensions such as sale price and size also show roughly a log-normal shape, though auctions sales seem to be more uniform over time (Figure 5). As shown in Figure 6, Impressionist pieces tend to be far more expensive than Contemporary pieces, which can be attributed to age. However, Contemporary pieces do tend to be physically larger (Figure 7) and have far more unbalanced dimensions, though as with Impressionist pieces large width and length do generally seem to scale together (Figure 8).

**RECENT ASSORTED ART (2006-2015)**

A major contribution of this research is the construction of a large dataset of recent auctions sales of assorted paintings[[4]](#footnote-4) (2006-2015), which was constructed by writing a computer program to scan recent listings on the Blouin Art Sales Index, a database that hosts a large collection of art auction data[[5]](#footnote-5). The motivation behind collecting and using a new dataset is twofold. First, the time gaps between auction sales in this dataset are on the much shorter scale of months, weeks, or even days, rather than years as in the previous datasets. This is far more conducive to studying anchoring. Second, this dataset consists of a very wide variety of artistic pieces, which is ideal for exploring substitutability across pieces. The Impressionist and Contemporary datasets tend to be more limited in their artistic scope, and so do not seem to be as conducive for studying substitution.

The raw dataset consists of approximately 500,000 observations, covering both 19th and 20th century art with some works from earlier time periods (earliest: approx. 1000 CE, for works by Song Dynasty artist Yi Yuanji). Nearly 90,000 artists are included, with the best represented being Pablo Picasso (1,868 works), Andy Warhol (1,712 works), and Joan Miro (880 works). However, the most expensive sale is an untitled crayon work by Cy Twombly (1928-2001), which went for $70.5 million at Sotheby’s in NYC in November 2015. The artists whose works sell for the most, on average, include Kazimir Malevich of the Suprematist movement (1879-1935), the Abstract Expressionist Mark Rothko (1903-1970), Vincent Van Gogh (1853-1890), and also Song Dynasty artists such as Emperor Huizong (1082-1135) and Yi Yuanji. As before, sale prices and other quantitative characteristics seem to follow a roughly log-normal distribution (Figure <>). Most of the paintings in this dataset sell for low 5 or 6 figure sums, while only a minority sell for higher figures reflected in the gradually diminishing right tail. It is clear that artists who sell more works through auction will enjoy higher revenue on average (regression slope: 0.52. p-val: <2E-16), as they become better-known in auction circles through higher representation (Figures 9-12).

Each observation in this new dataset includes the artwork title, the artist, artwork category as described by the auction house, a textual description of the materials, the lot number, sale date, auction house, and the USD sale price. Because information on the materials were given in the form of unstructured text data, which might be attributed to freeform data entry by Blouin, simple keyword extraction was used to extract hedonic characteristics such as height and width; more sophisticated textual extraction methods should be employed in future work. Some summary statistics for the full raw dataset are provided in Table 3.

1. Richardson (2002); Abowd & Ashenfelter (1989); Beggs & Graddy (1997); Ashenfelter & Graddy (2003); Beggs & Graddy (2009) [↑](#footnote-ref-1)
2. http://www.jstor.org/stable/pdf/2556028.pdf?acceptTC=true [↑](#footnote-ref-2)
3. http://people.brandeis.edu/~kgraddy/data.html [↑](#footnote-ref-3)
4. The collected raw data also includes prints, drawings, and other mediums, but since we wish to compare against our other two datasets, we only use paintings here. [↑](#footnote-ref-4)
5. http://artsalesindex.artinfo.com [↑](#footnote-ref-5)