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] and are already described in detail elsewhere[2]. Both of those datasets are available on the Brandeis website of Kathryn Graddy[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

---

[1] Richardson (2002); Abowd & Ashenfelter (1989); Beggs & Graddy (1997); Ashenfelter & Graddy (2003); Beggs & Graddy (2009)
[2] http://www.jstor.org/stable/pdf/2556028.pdf?acceptTC=true
[3] http://people.brandeis.edu/~kgraddy/data.html

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.

Table <>: Impressionist art, summary statistics for continuous features.

```
      DIM_A               LOW_EST              HIGH_EST
Min.   :  0.00    Min.   :     102    Min.   :     128
1st Qu.: 11.00    1st Qu.:   14000    1st Qu.:   18000
Median : 17.00    Median :   40000    Median :   50000
Mean   : 18.31    Mean   :  196023    Mean   :  257967
3rd Qu.: 23.00    3rd Qu.:  132800    3rd Qu.:  168300
Max.   :120.00    Max.   :40000000    Max.   :50000000
                  NA's   :37

     S_PRICE            CNV_RATE            DATE_PTG
 Min.   :     126   Min.   :0.0000    Min.   :1823
 1st Qu.:   18700   1st Qu.:0.0000    1st Qu.:1902
 Median :   53856   Median :1.2400    Median :1922
 Mean   :  285428   Mean   :0.8639    Mean   :1921
 3rd Qu.:  176000   3rd Qu.:1.6800    3rd Qu.:1938
 Max.   :82500000   Max.   :2.3610    Max.   :1983
 NA's   :4696                         NA's   :3950

     DATE_FLG            DIM_B                DIAM
 Min.   :0.0000    Min.   :  0.00    Min.   : 1.00
 1st Qu.:0.0000    1st Qu.: 11.00    1st Qu.: 6.75
 Median :0.0000    Median : 18.00    Median :11.50
 Mean   :0.3538    Mean   : 18.69    Mean   :15.10
 3rd Qu.:1.0000    3rd Qu.: 24.00    3rd Qu.:24.50
 Max.   :1.0000    Max.   :141.00    Max.   :36.00
         NA's   :37        NA's   :16243

                      PND_FLG
                 Min.   :0.0000
                 1st Qu.:0.0000
                 Median :1.0000
                 Mean   :0.5127
                 3rd Qu.:1.0000
                 Max.   :2.0000
                   NA's   :4
```

Table < > shows summary statistics for continuous (non-categorical) variables in the Impressionist art dataset. 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. Additionally, many of these quantities have distributions that are roughly log-normal (i.e., without the log transformation, skewed heavily right), showed in Figure <>.
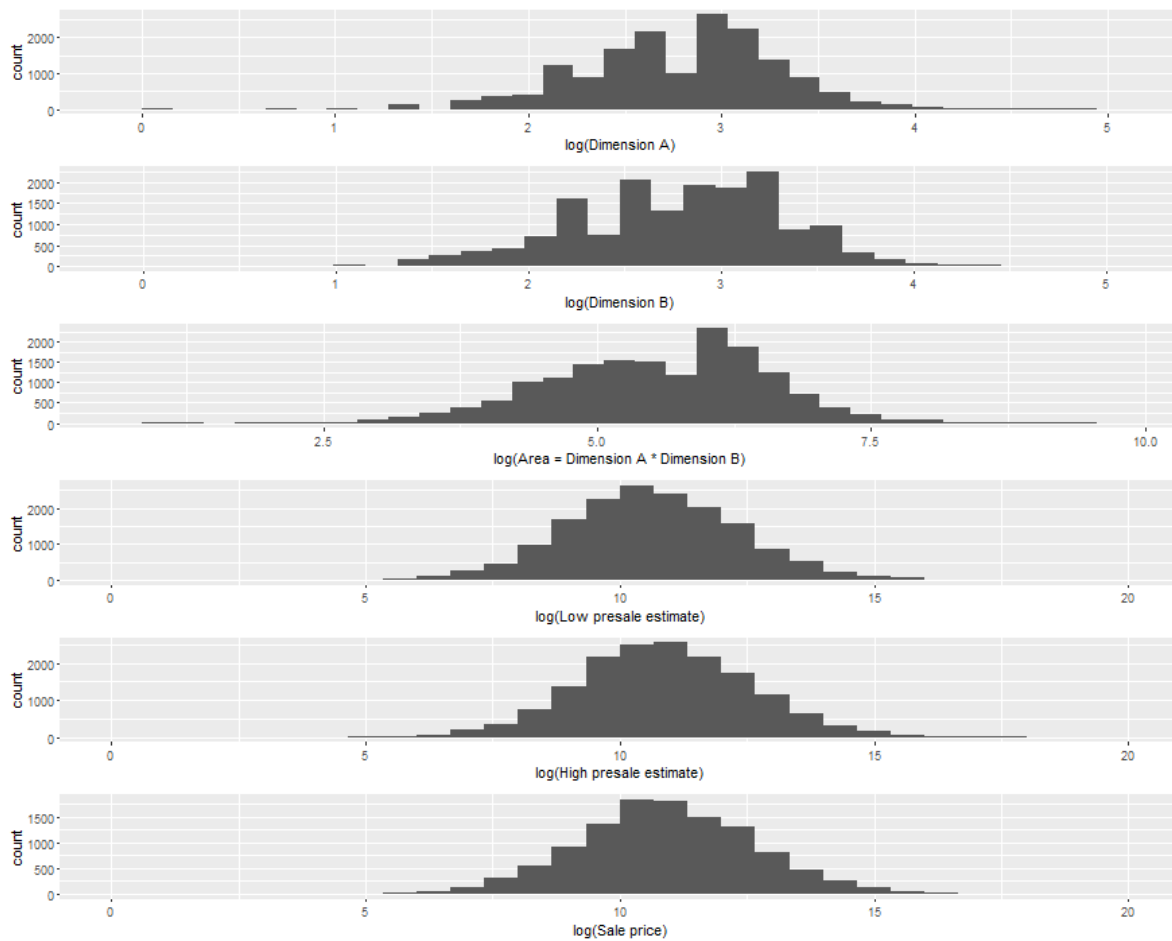
This is because the majority of paintings exhibit middle-market sale price, estimates, size, and so forth, while relatively few reach the highest ranges. The two painting dimensions have the most irregular distributions, particularly in the middle ranges.

However, as seen in Figure ◇, 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, though there are clearly some huge years with record numbers of sales.

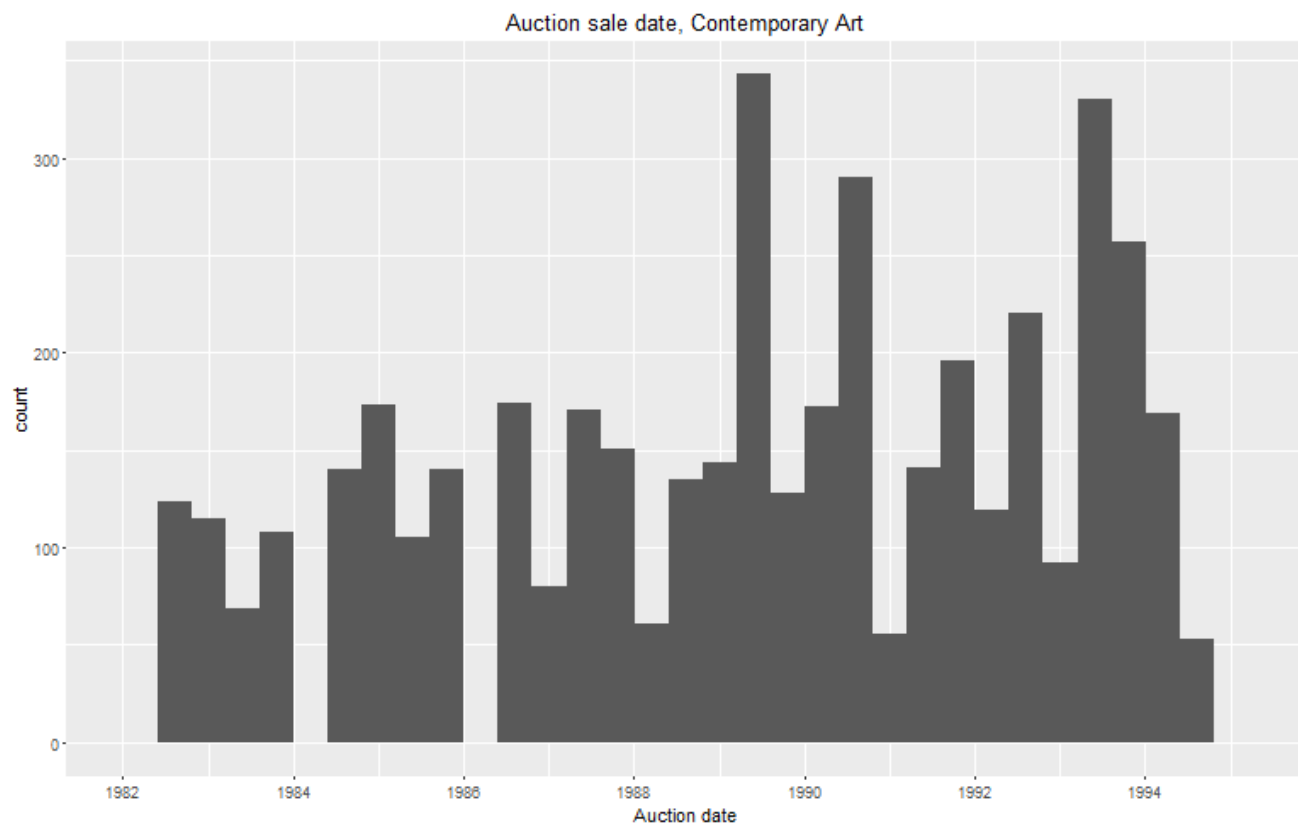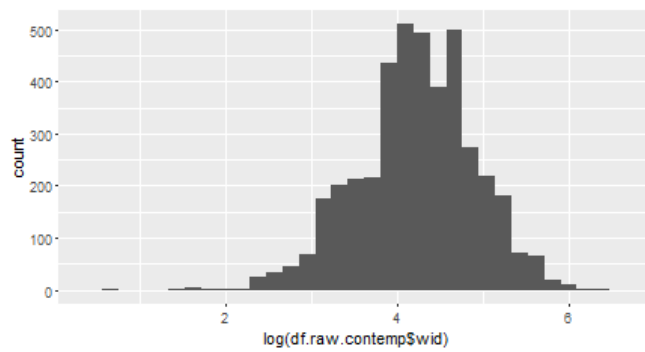Auction sale dates, Impressionist Art

Comparison of painting dimensions, Impressionist Art
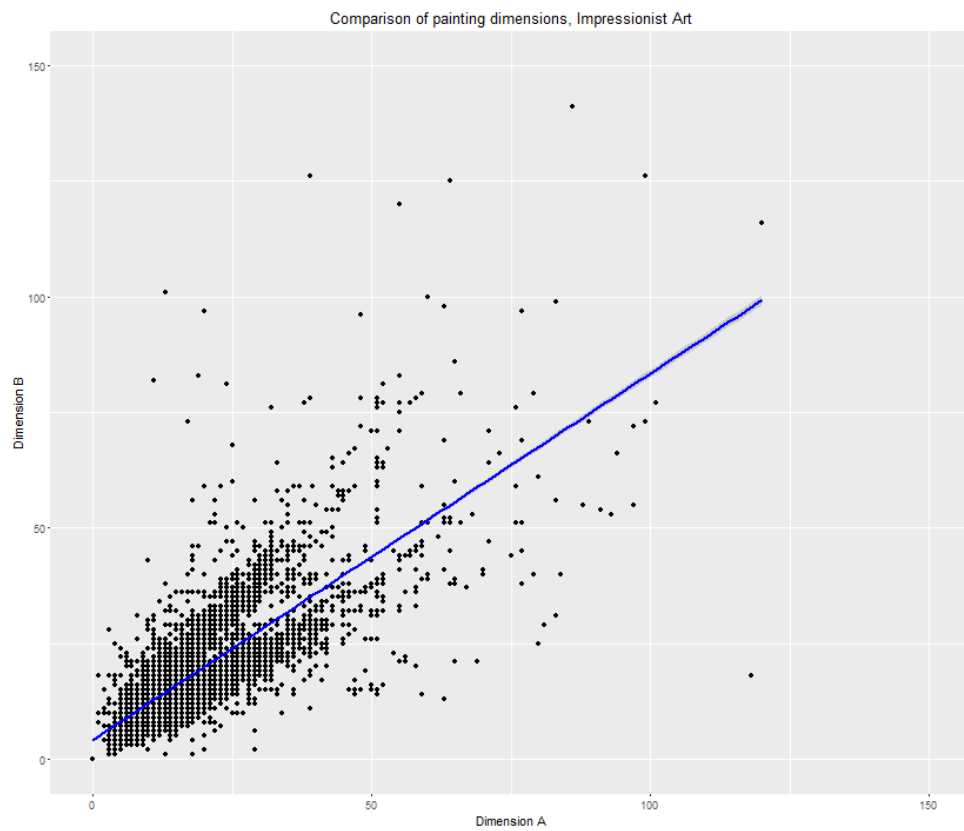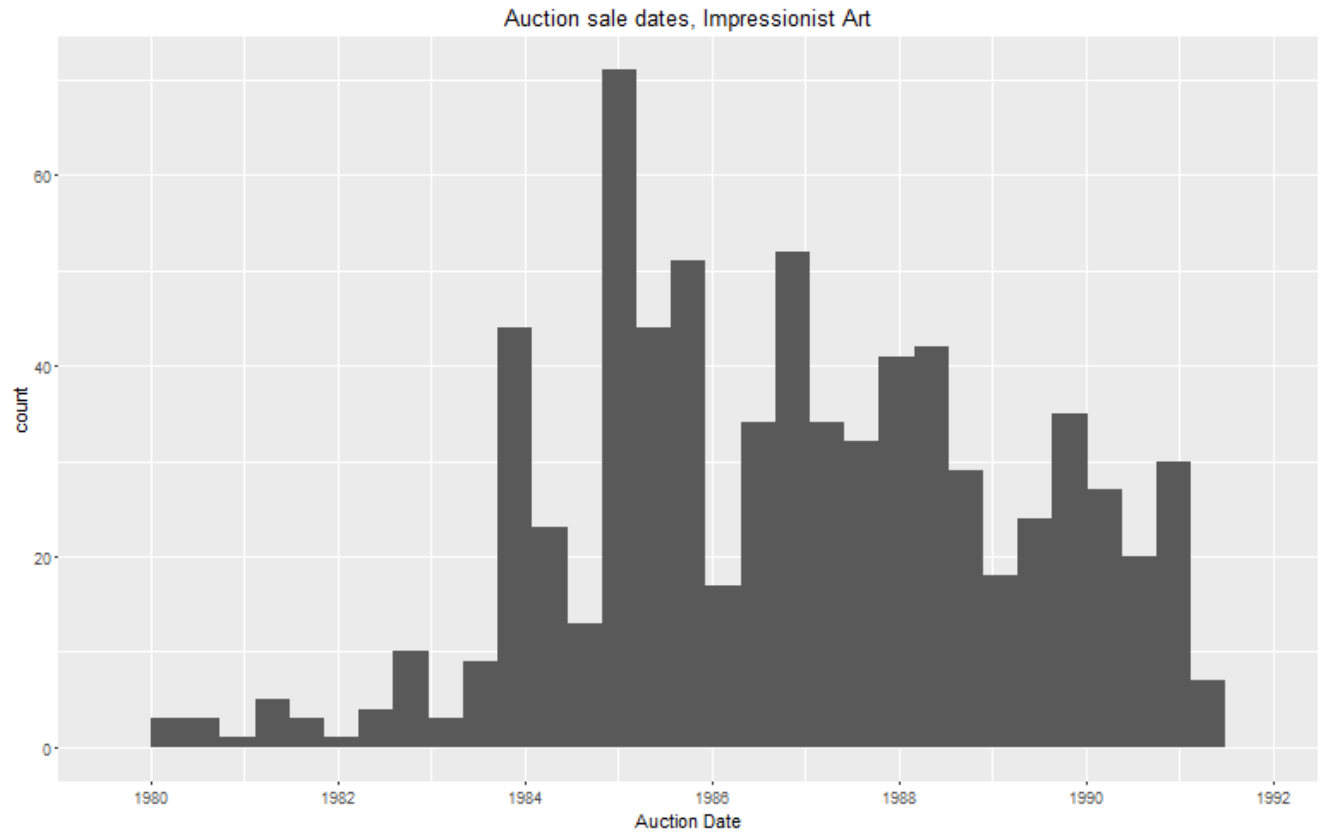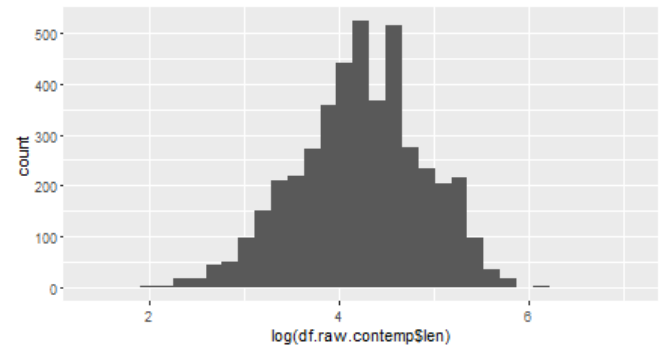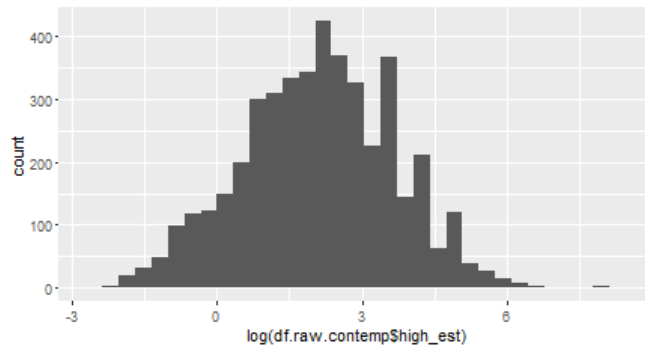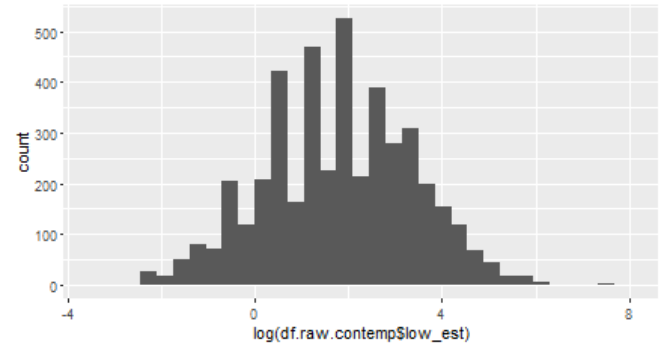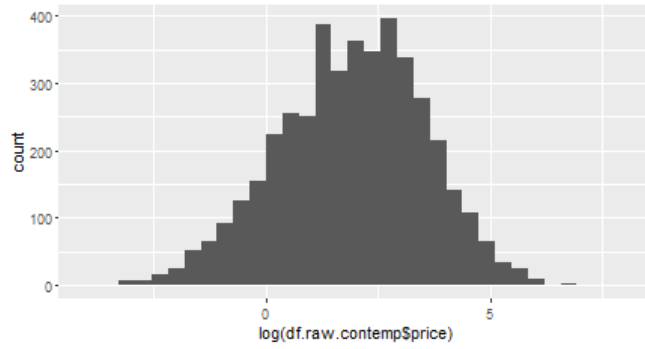
CONTEMPORARY ART (1982-1994)

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.
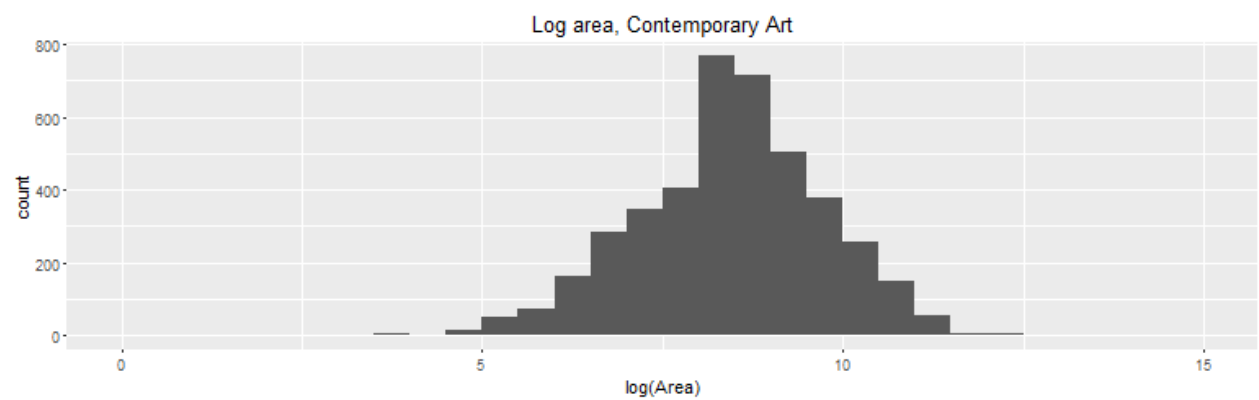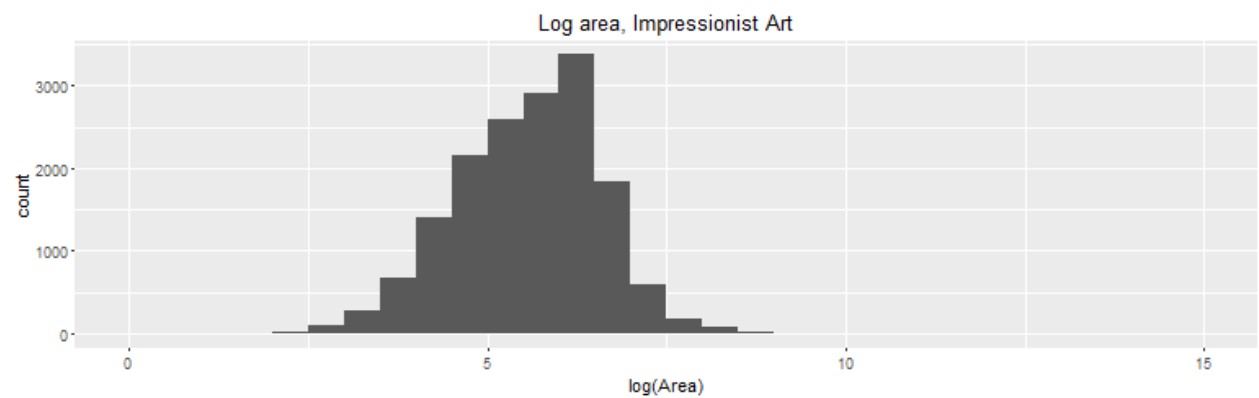
Table 2: Contemporary art, summary statistics.

```
   Auction_date            mdate              ddate               ydate
 Min.   :1982-06-29  Min.   : 2.000   Min.   : 1.00   Min.   :1982
 1st Qu.:1986-06-26  1st Qu.: 6.000   1st Qu.: 5.00   1st Qu.:1986
 Median :1989-06-29  Median : 6.000   Median :22.00   Median :1989
 Mean   :1989-05-15  Mean   : 7.831   Mean   :17.07   Mean   :1989
 3rd Qu.:1992-07-02  3rd Qu.:12.000   3rd Qu.:26.00   3rd Qu.:1992
 Max.   :1994-06-30  Max.   :12.000   Max.   :30.00   Max.   :1994

      lot               sold            price             low_est
 Min.   :   1.0   Min.   :0.0000   Min.   :  0.00   Min.   :  0.05
 1st Qu.:  87.0   1st Qu.:1.0000   1st Qu.:  1.90   1st Qu.:  2.00
 Median : 423.0   Median :1.0000   Median :  7.00   Median :  6.00
 Mean   : 397.7   Mean   :0.7745   Mean   : 21.23   Mean   : 19.53
```

```
3rd Qu.: 601.0   3rd Qu.:1.0000   3rd Qu.:  20.00   3rd Qu.:  20.00
Max.   :1164.0   Max.   :1.0000   Max.   :1700.00   Max.   :1800.00
                 NA's   :2        NA's   :45
    high_est          date_ptg           len               wid
 Min.   :   0.1   Min.   :26.00    Min.   :  5.40    Min.   :  2.00
 1st Qu.:   3.0   1st Qu.:60.00    1st Qu.: 44.50    1st Qu.: 46.00
 Median :   8.0   Median :67.00    Median : 70.00    Median : 70.00
 Mean   :  26.1   Mean   :68.24    Mean   : 84.53    Mean   : 84.71
 3rd Qu.:  25.0   3rd Qu.:77.00    3rd Qu.:105.00    3rd Qu.:105.00
 Max.   :2600.0   Max.   :91.00    Max.   :957.00    Max.   :602.00
  NA's   :45       NA's   :449      NA's   :73        NA's   :293
    artist            medium           CNV_RATE          ukcpi
 Length:4456      Length:4456       Min.   :1.210    Min.   :239.6
 Class :character Class :character  1st Qu.:1.482    1st Qu.:286.4
 Mode  :character Mode  :character  Median :1.610    Median :339.3
                                    Mean   :1.609   Mean   :342.9
                                    3rd Qu.:1.722   3rd Qu.:407.1
                                    Max.   :1.954   Max.   :423.0

     ukinf             uktb             uscpi             usinf
 Min.   : 1.270   Min.   : 4.900   Min.   :181.6    Min.   :1.280
 1st Qu.: 3.050   1st Qu.: 8.800   1st Qu.:204.1    1st Qu.:3.050
 Median : 4.710   Median : 9.630   Median :231.7    Median :3.920
 Mean   : 5.061   Mean   : 9.832   Mean   :232.7    Mean   :3.848
 3rd Qu.: 6.520   3rd Qu.:11.990   3rd Qu.:261.9    3rd Qu.:4.600
 Max.   :10.430   Max.   :14.540   Max.   :276.8    Max.   :6.220

     ustb             japcpi             dj               ftse
 Min.   : 2.970   Min.   :149.3    Min.   : 812.2   Min.   : 736.2
 1st Qu.: 3.990   1st Qu.:160.6    1st Qu.:1776.5   1st Qu.:1588.4
 Median : 6.990   Median :168.2    Median :2458.3   Median :2182.0
 Mean   : 6.157   Mean   :169.9    Mean   :2438.5   Mean   :2078.3
 3rd Qu.: 7.760   3rd Qu.:182.3    3rd Qu.:3174.7   3rd Qu.:2546.6
 Max.   :10.320   Max.   :185.4    Max.   :3753.5   Max.   :3223.9

                      VAT
                 Min.   :0.0000
                 1st Qu.:0.0000
                 Median :0.0000
                 Mean   :0.2949
                 3rd Qu.:1.0000
                 Max.   :1.0000
```
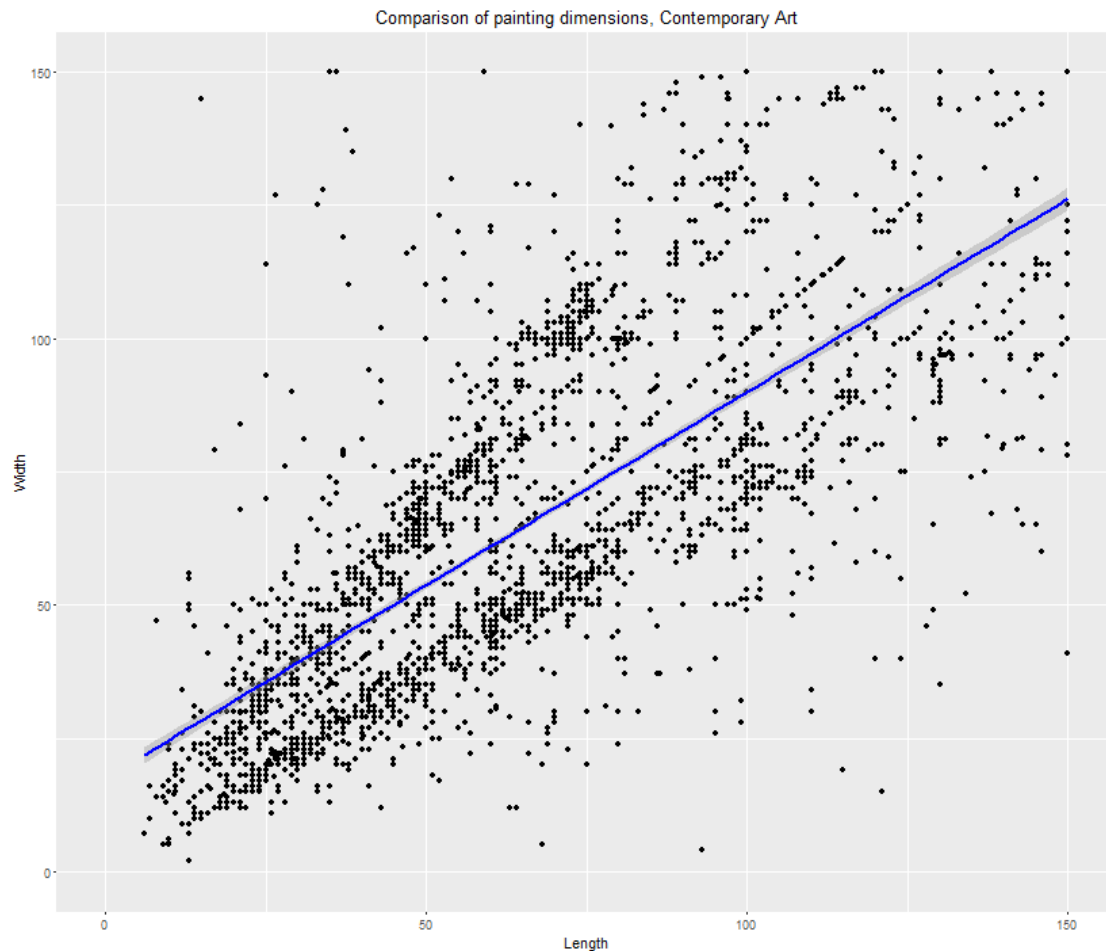
Auction sale date, Contemporary Art

USD log prices, Impressionist Art

USD log prices, Contemporary Art

Log area, Impressionist Art

Log area, Contemporary Art

Comparison of painting dimensions, Contemporary Art
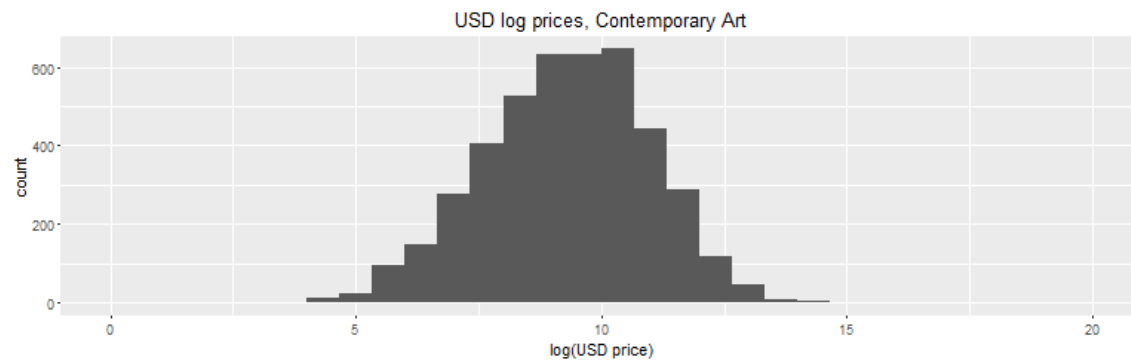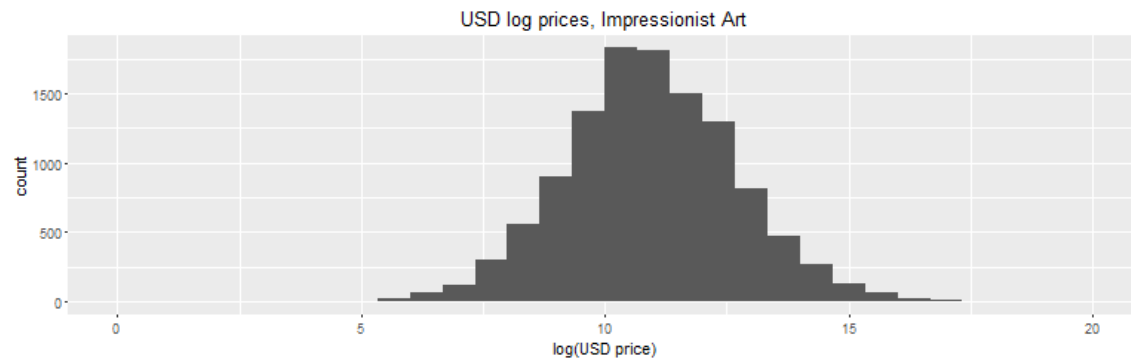
Table <> gives summary statistics for this dataset. As with the Impressionist dataset,

and shown in Figure <>, 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 <>). As shown in Figure <>, Impressionist pieces tend to be far more expensive

than Contemporary pieces, which can be attributed to age effects. However,

Contemporary pieces do tend to be physically larger (Figure <>) and have far more

unbalanced dimensions, though as with Impressionist pieces large width and length do

generally seem to scale together (Figure <>).
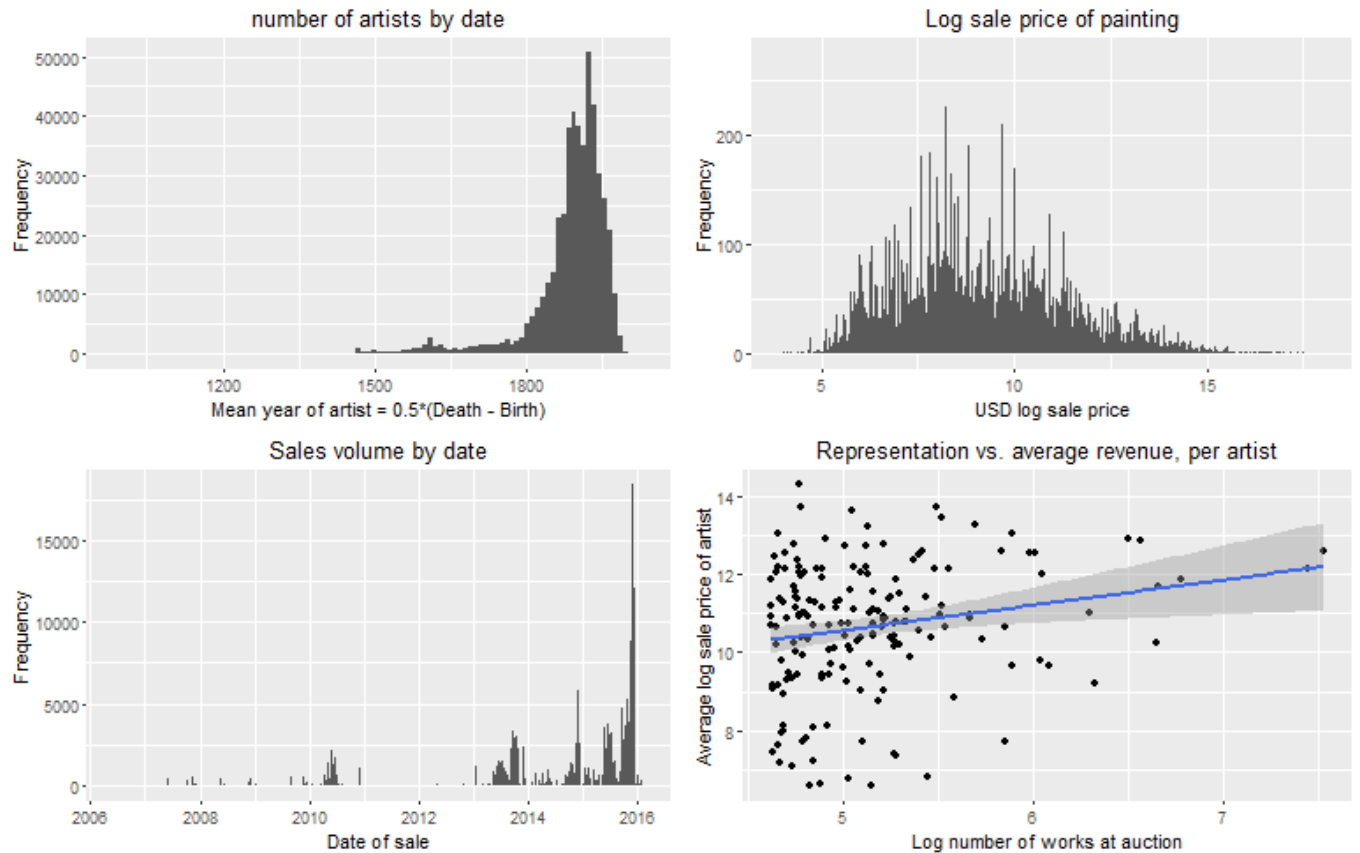
RECENT ASSORTED ART (2006-2015)

A major contribution of this research is the construction of a large dataset of recent auctions sales of assorted paintings[4] (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]. 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

---

[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.

[5] http://artsalesindex.artinfo.com

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 (Figure <>).

number of artists by date

Log sale price of painting

Sales volume by date

Representation vs. average revenue, per artist

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.

Table 3: Assorted art, summary statistics.

```
     height            width          area.inches        artist.startdate
 Min.   :      0   Min.   :       0   Min.   :0.000e+00   Min.   :1000
 1st Qu.:     12   1st Qu.:      12   1st Qu.:1.520e+02   1st Qu.:1869
 Median :     19   Median :      20   Median :3.920e+02   Median :1904
 Mean   :     64   Mean   :      78   Mean   :2.270e+08   Mean   :1886
 3rd Qu.:     29   3rd Qu.:      29   3rd Qu.:8.160e+02   3rd Qu.:1932
 Max.   :7700281   Max.   :10197670   Max.   :7.852e+13   Max.   :2015
 NA's   :4000      NA's   :31325      NA's   :86729       NA's   :19411


 artist.enddate    lot.number        sale.date          usd.sale.price
 Min.   :1016   Min.   :     0   Min.   :2006-06-09   Min.   :       1
 1st Qu.:1930   1st Qu.:    81   1st Qu.:2013-10-15   1st Qu.:     905
 Median :1956   Median :   205   Median :2015-06-02   Median :    3009
 Mean   :1941   Mean   :  1195   Mean   :2014-08-27   Mean   :   50275
 3rd Qu.:1983   3rd Qu.:   599   3rd Qu.:2015-11-11   3rd Qu.:   12188
 Max.   :2015   Max.   :221186   Max.   :2016-02-04   Max.   :70530000
 NA's   :19411  NA's   :275      NA's   :275          NA's   :209591
```