What Makes Art Valuable: Data Scraping and Exploratory Data Visualizations

By Georgie Coetzer, Marcus Ribeiro, Ramiro Storni

Using Auction Data to Explore Art Market Trends

Project Overview

Art is often a polarizing topic of discussion. Whether you love it or hate it, the price for art can reach exorbitant levels, especially in luxury auction settings.

Works of art routinely sell for hundreds of thousands of dollars. In some instances, artworks like Leonardo Da Vinci's *Salvator Mundi* can even sell for hundreds of millions of dollars. With this in mind, we sought to investigate what factors might influence the price of artworks at auction.

Our hypothesis was that the prestige, or the fame/popularity, of an artist would be one of the most significant factors in influencing the price of an artwork. Moreover, the article *Are Art Auction Estimates Biased?* leads us to believe that auction house estimates are an unreliable metric, generally undervaluing artworks.

In our efforts to answer this question, we took the following steps:

1. Data Collection
2. Data Cleaning
3. Data Processing

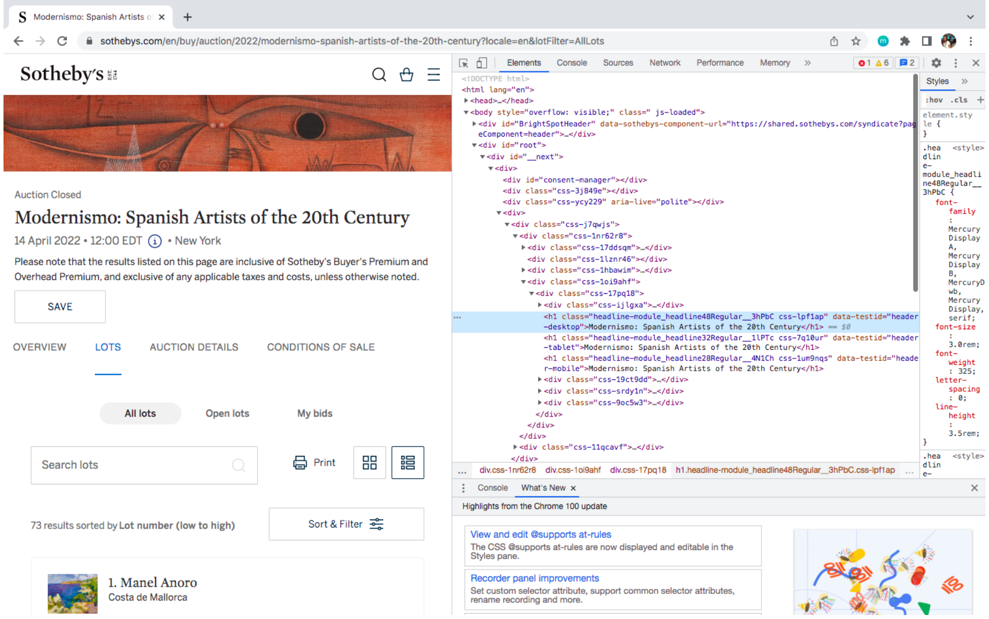
4. Data visualization

Data Collection

To gather our data, we used web scrapers we constructed using Selenium, a Python library and tool for automating web browsers to perform a variety of tasks. We chose Selenium because the websites we are scraping data from – Sotheby's and Christie's– are dynamic and Selenium provides an easy-to-understand interface that interacts with websites in a similar manner to the way in which we as humans interact with websites. Selenium, therefore, enabled us to interact with these dynamic websites. Building the web scrapers involved inspecting the web pages to locate the Xpath associated with the relevant filters we wanted to apply and data we wanted to collect and then using Selenium functions to interact with and access these elements

The functionality of the scrapers can be divided into three parts:

1. Collecting the URLs for each auction—We filtered the Auction results to see only Fine Art and Prints/Photographs
2. Collect the URLs for each artwork in an auction
3. Gather key features, such as price, artists, estimate price, etc. for each artwork



Auction Closed

Modernismo: Spanish Artists of the 20th Century

14 April 2022 • 12:00 EDT • New York

Please note that the results listed on this page are inclusive of Sotheby's Buyer's Premium and Overhead Premium, and exclusive of any applicable taxes and costs, unless otherwise noted.

73 results sorted by **Lot number (low to high)**

1. Manel Anoro
Costa de Mallorca

```
<!DOCTYPE html>
<html lang="en">
 <head>…</head>
 <body style="overflow: visible;" class=" js-loaded">
  <div id="BrightSpotHeader" data-sothebys-component-url="https://shared.sothebys.com/syndicate?pag
  eComponent=header">…</div>
  <div id="root">
   <div id="__next">
    <div>
     <div id="consent-manager"></div>
     <div class="css-3j849e"></div>
     <div class="css-ycy229" aria-live="polite"></div>
     <div>
      <div class="css-j7qwjs">
       <div class="css-1nr62r8">
        <div class="css-17ddsqm">…</div>
        <div class="css-1lznr46"></div>
        <div class="css-1hbawim">…</div>
        <div class="css-1oi9ahf">
         <div class="css-17pq18">
          <div class="css-ijlgxa">…</div>
          <h1 class="headline-module_headline48Regular__3hPbC css-lpf1ap" data-testid="header
          -desktop">Modernismo: Spanish Artists of the 20th Century</h1> == $0
          <h1 class="headline-module_headline32Regular__1lPTc css-7q10ur" data-testid="header
          -tablet">Modernismo: Spanish Artists of the 20th Century</h1>
          <h1 class="headline-module_headline28Regular__4N1Ch css-1um9nqs" data-testid="heade
          r-mobile">Modernismo: Spanish Artists of the 20th Century</h1>
          <div class="css-19ct9dd">…</div>
          <div class="css-srdy1n">…</div>
          <div class="css-9oc5w3">…</div>
         </div>
        </div>
       </div>
       <div class="css-11qcavf">…</div>
      </div>
... div.css-1nr62r8 div.css-1oi9ahf div.css-17pq18 h1.headline-module_headline48Regular__3hPbC.css-lpf1ap ...
```

```
element.style {
}
.headline-module_headline48Regular__3hPbC {   <style>
  font-family: Mercury Display A, Mercury Display B, MercuryDwb, Mercury Display, serif;
  font-size: 3.0rem;
  font-weight: 325;
  letter-spacing: 0;
  line-height: 3.5rem;
}
.headline-   <style>
```

Highlights from the Chrome 100 update

View and edit @supports at-rules
The CSS @supports at-rules are now displayed and editable in the Styles pane.

Recorder panel improvements
Set custom selector attribute, support common selector attributes, rename recording and more.

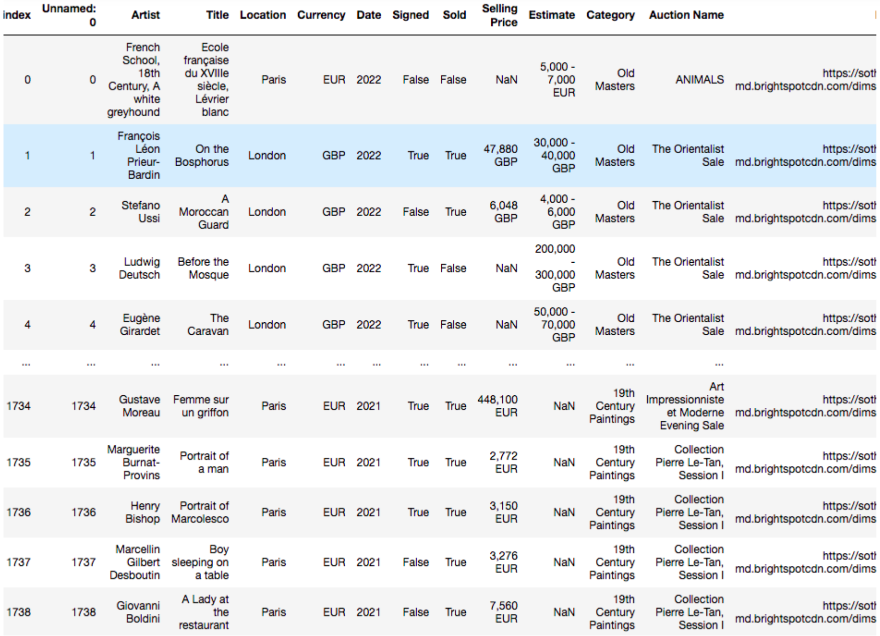
Fig 1. Inspecting web elements for a Sotheby's Auction

Ultimately, we were able to scrape data for 1738 artworks from Sotheby's auctions and 4525 from Christie's.

Data Cleaning:

The data we collected required significant cleaning before we would be able to use it. As part of the data cleaning process, we had to separate the high and low estimate prices and convert all the prices to a single currency – which we chose to be USD– because the artworks were sold in different countries so the sale currency differed. Similarly, numeric values such as prices and estimates were initially stored as strings in the dataframes, and names were sometimes formatted differently. Moreover, we had to filter out unwanted art forms such as furniture pieces and sculptures which our scrapers had failed to filter out.

For our data cleaning process, we used pandas data frames and ran a variety of cleaning functions that would strip and modify strings, create new columns, and ultimately enable us to work with the data. As part of this process, we chose to drop certain artworks which lacked key features.



index	Unnamed: 0	Artist	Title	Location	Currency	Date	Signed	Sold	Selling Price	Estimate	Category	Auction Name	
0	0	French School, 18th Century, A white greyhound	Ecole française du XVIIIe siècle, Lévrier blanc	Paris	EUR	2022	False	False	NaN	5,000 - 7,000 EUR	Old Masters	ANIMALS	https://soth md.brightspotcdn.com/dims
1	1	François Léon Prieur-Bardin	On the Bosphorus	London	GBP	2022	True	True	47,880 GBP	30,000 - 40,000 GBP	Old Masters	The Orientalist Sale	https://soth md.brightspotcdn.com/dims
2	2	Stefano Ussi	A Moroccan Guard	London	GBP	2022	False	True	6,048 GBP	4,000 - 6,000 GBP	Old Masters	The Orientalist Sale	https://soth md.brightspotcdn.com/dims
3	3	Ludwig Deutsch	Before the Mosque	London	GBP	2022	True	False	NaN	200,000 - 300,000 GBP	Old Masters	The Orientalist Sale	https://soth md.brightspotcdn.com/dims
4	4	Eugène Girardet	The Caravan	London	GBP	2022	True	False	NaN	50,000 - 70,000 GBP	Old Masters	The Orientalist Sale	https://soth md.brightspotcdn.com/dims
...	...	...	...	...	...	...	...	...	...	...	...	...	
1734	1734	Gustave Moreau	Femme sur un griffon	Paris	EUR	2021	True	True	448,100 EUR	NaN	19th Century Paintings	Art Impressionniste et Moderne Evening Sale	https://soth md.brightspotcdn.com/dims
1735	1735	Marguerite Burnat-Provins	Portrait of a man	Paris	EUR	2021	True	True	2,772 EUR	NaN	19th Century Paintings	Collection Pierre Le-Tan, Session I	https://soth md.brightspotcdn.com/dims
1736	1736	Henry Bishop	Portrait of Marcolesco	Paris	EUR	2021	True	True	3,150 EUR	NaN	19th Century Paintings	Collection Pierre Le-Tan, Session I	https://soth md.brightspotcdn.com/dims
1737	1737	Marcellin Gilbert Desboutin	Boy sleeping on a table	Paris	EUR	2021	False	True	3,276 EUR	NaN	19th Century Paintings	Collection Pierre Le-Tan, Session I	https://soth md.brightspotcdn.com/dims
1738	1738	Giovanni Boldini	A Lady at the restaurant	Paris	EUR	2021	False	True	7,560 EUR	NaN	19th Century Paintings	Collection Pierre Le-Tan, Session I	https://soth md.brightspotcdn.com/dims

Fig 2. Initial dataframe for Sotheby's

Data Processing:

Once we cleaned our data, we chose to add features that we hoped would give us some insight into the pricing of an artwork at auction. Using Pandas Data Frames in conjunction with the beautiful soup and requests libraries as well, as the Yahoo API, we collected the number of Yahoo search results for each artist in our dataset. We interpreted this value as an artist's popularity and believed that if an artist had more search results we might see an increase in an artwork's price.

We also worked with the existing data to create new features such as artist age, whether an artist is alive, whether or not an artwork sold, and whether or not the auction house accurately estimated the price of the artwork.

Visual Features

In addition to the features we were able to extract from the auction sites, we wanted to include visual features in our analysis. We followed Jason Brownlee's tutorial to implement a pre-trained image classifier and planned on using the pre-trained VGG16 convolutional neural network to add a visual feature to our data by adding what the network predicts to be in the artwork with the highest probability. However, after some initial tests, we, ultimately, did not include this visual feature in our analysis as the network failed to classify images that we thought would be fairly straightforward such as portraits. It was entertaining, however, to see how the network classified abstract art.

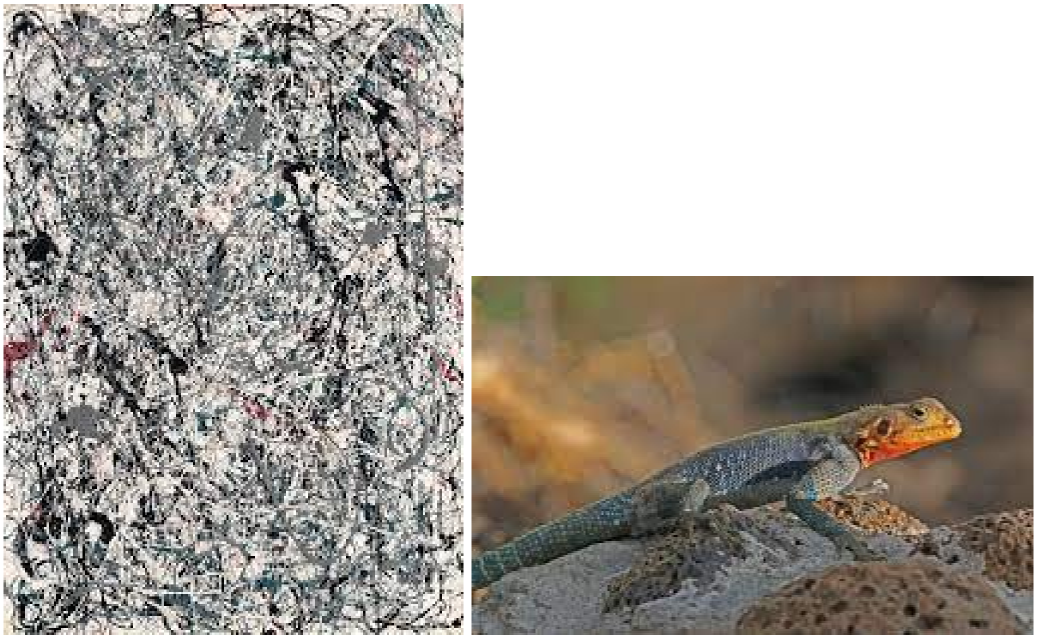


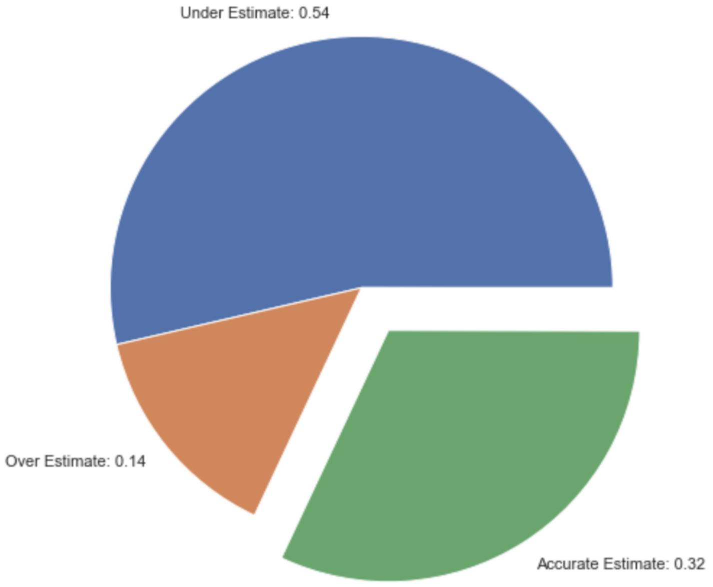
Fig 3. Network classified Jackson Pollock's Number 19 (left) as an Agama Lizard (right)

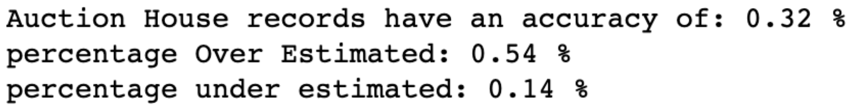
Data Exploration and Visualization

Once we had completed the required data cleaning and processing, we chose to focus our analysis on the data scraped from Christie’s as we struggled to get the Sotheby’s data into a workable format as data for many of the artworks lacked key features.

Now that we had data, we ran a few exploratory visualizations:

When we first started researching auction houses for this project, we read that auction houses tended to underestimate the value of their works of art to make the house look better when artworks sold over the estimated price. Here we visualized the percentage of works that sold under, within, and over the estimated prices in a pie chart:



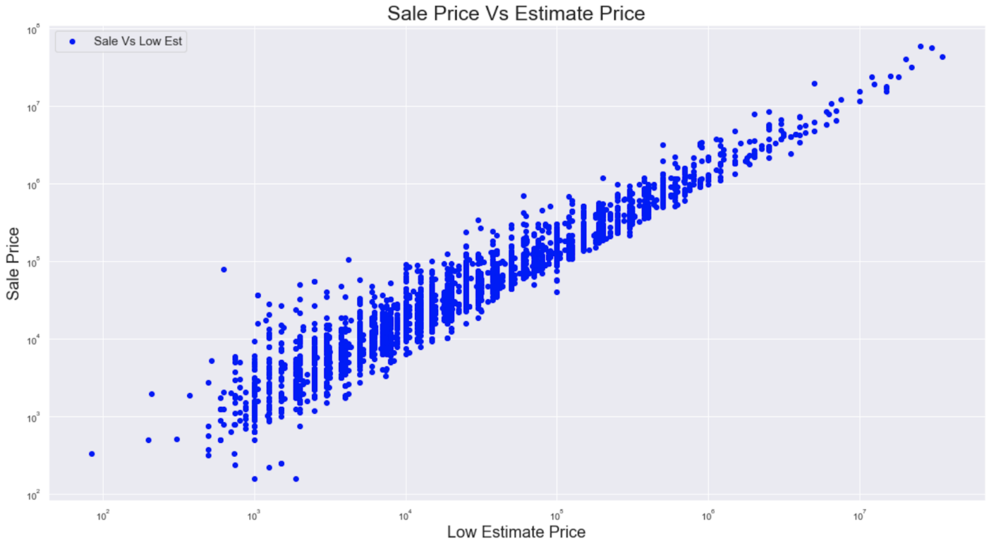


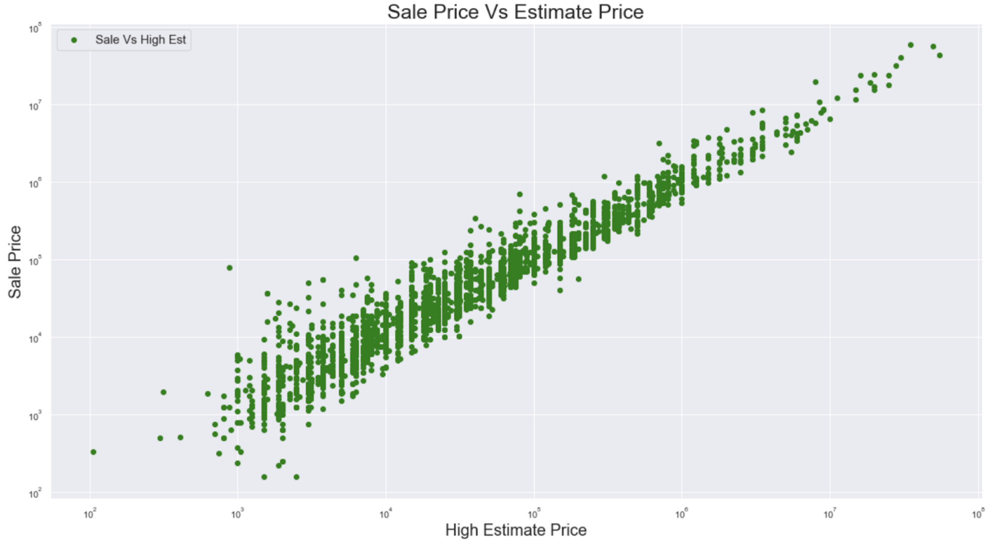
```
Auction House records have an accuracy of: 0.32 %
percentage Over Estimated: 0.54 %
percentage under estimated: 0.14 %
```

This data seems to prove the statement that auction houses tend to underestimate the values of artworks by showing that they underestimate approximately 54% of the time.

Correlations:

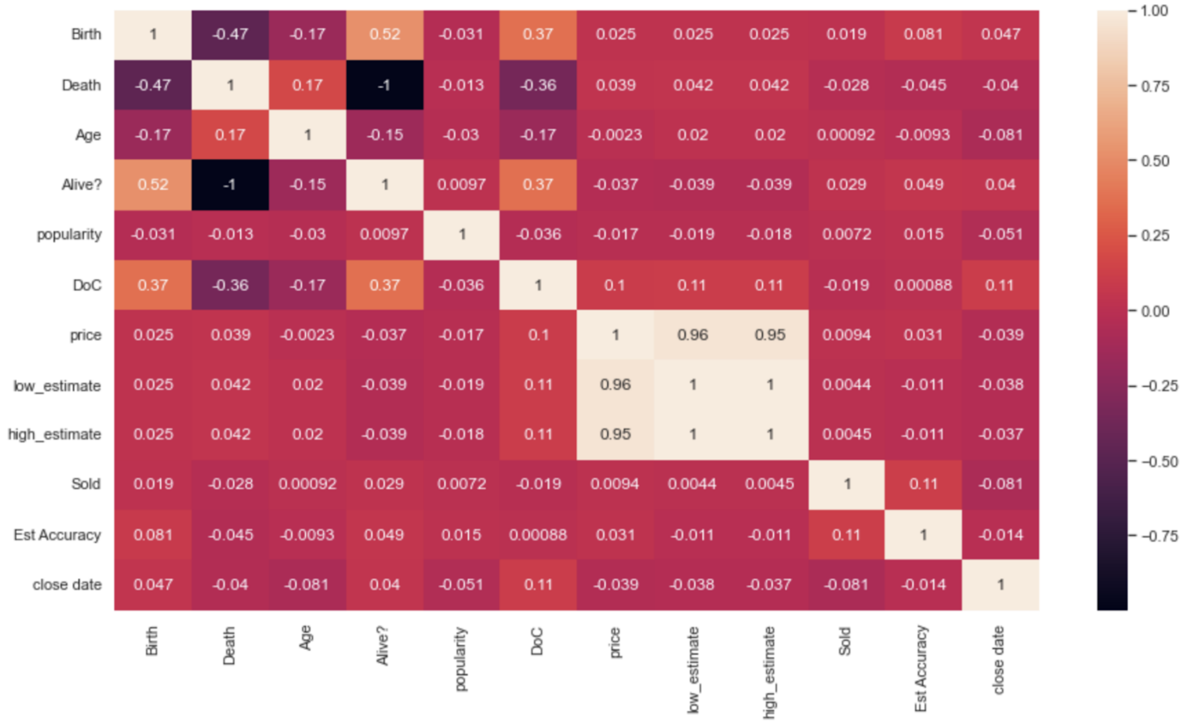
Upon seeing this we wondered how closely the sale price of a work of art and the estimated prices correlated to the estimated prices given by the auction house.





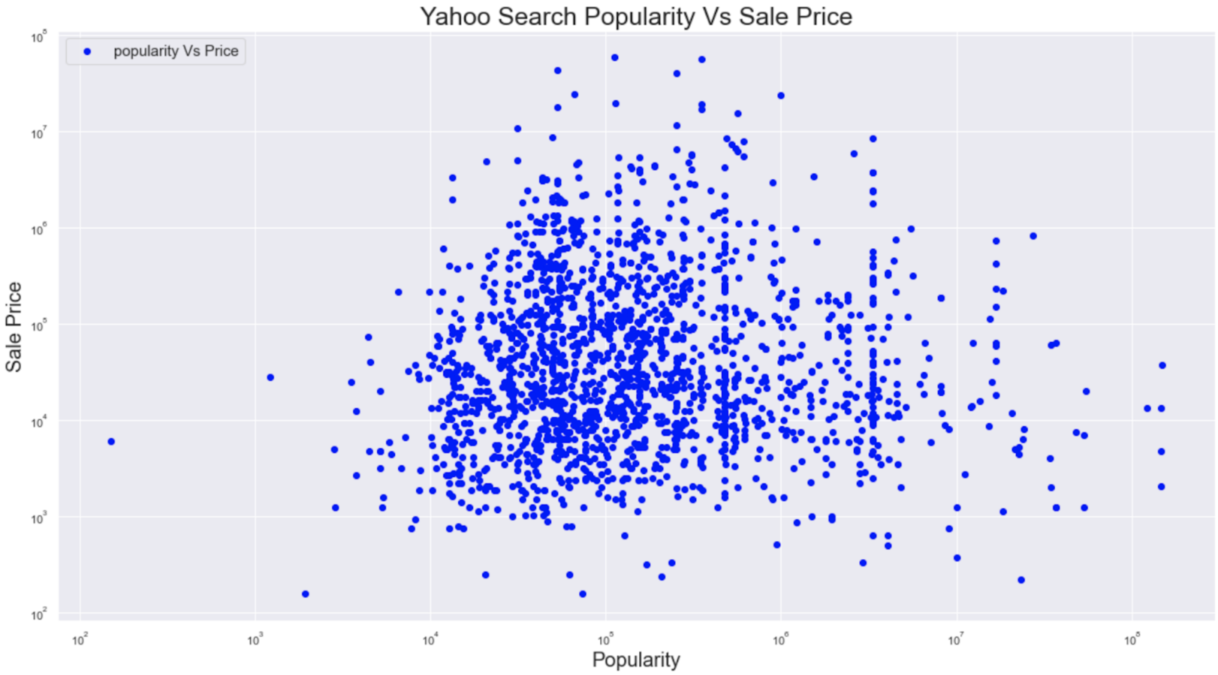
The above scatter plots show a high correlation between the sale price of an artwork and the estimates given by the auction house prior to the auction. We concluded that this was likely a result of the anchoring bias that occurs when the auction house gives its evaluation of a work of art at the beginning of the auction. If this is the case, the data seems to imply that the auction house estimates which are set by professional art appraisers are one of the primary influences on the value assigned to an artwork.

This is further reinforced by our confusion matrix which shows very little correlation between any of the features except for price and given estimates.



	Birth	Death	Age	Alive?	popularity	DoC	price	low_estimate	high_estimate	Sold	Est Accuracy	close date
Birth	1	-0.47	-0.17	0.52	-0.031	0.37	0.025	0.025	0.025	0.019	0.081	0.047
Death	-0.47	1	0.17	-1	-0.013	-0.36	0.039	0.042	0.042	-0.028	-0.045	-0.04
Age	-0.17	0.17	1	-0.15	-0.03	-0.17	-0.0023	0.02	0.02	0.00092	-0.0093	-0.081
Alive?	0.52	-1	-0.15	1	0.0097	0.37	-0.037	-0.039	-0.039	0.029	0.049	0.04
popularity	-0.031	-0.013	-0.03	0.0097	1	-0.036	-0.017	-0.019	-0.018	0.0072	0.015	-0.051
DoC	0.37	-0.36	-0.17	0.37	-0.036	1	0.1	0.11	0.11	-0.019	0.00088	0.11
price	0.025	0.039	-0.0023	-0.037	-0.017	0.1	1	0.96	0.95	0.0094	0.031	-0.039
low_estimate	0.025	0.042	0.02	-0.039	-0.019	0.11	0.96	1	1	0.0044	-0.011	-0.038
high_estimate	0.025	0.042	0.02	-0.039	-0.018	0.11	0.95	1	1	0.0045	-0.011	-0.037
Sold	0.019	-0.028	0.00092	0.029	0.0072	-0.019	0.0094	0.0044	0.0045	1	0.11	-0.081
Est Accuracy	0.081	-0.045	-0.0093	0.049	0.015	0.00088	0.031	-0.011	-0.011	0.11	1	-0.014
close date	0.047	-0.04	-0.081	0.04	-0.051	0.11	-0.039	-0.038	-0.037	-0.081	-0.014	1

Surprisingly, the popularity of an artist—which we measured using Yahoo search results—had no correlation to the sale price of a work of art at auction.



Challenges:

Our biggest challenge in this project was data collection. Prior to this project, none of us had ever worked with Selenium before and there was a learning curve that we had to get over before we were able to get started. Even after we understood how to use the library, a lack of uniformity in the auction and artwork sites made it difficult to scrape the same features across auctions and artworks. For the most part, we were able to overcome this using error handling but it slowed us down significantly.

The challenges of scraping features consistently across web pages made it difficult for us to narrow our scope to one type of artwork such as paintings or prints, as it was difficult to filter the data we were collecting. This was made worse by the fact that the structure of the web pages prevented us from including medium as a feature. Unfortunately, this issue of scope might have also introduced noise to the data. For example, a print by Picasso will sell for much less than one of his paintings but we do not take that into consideration when comparing the price of an artwork with the artist's popularity.

Conclusion:

Through our analysis of Christie's and Sotheby's auction data, it appears as though the most significant factor influencing the price an artwork will sell for at auction, is the estimates provided by the auction house. Based on this finding, we believe that there may be an anchoring bias that is influencing the buyer's decision-making. Although the estimates were not wholly accurate in predicting prices it seems as though the selling prices of artworks at auction will be closely correlated to the given auction house estimates.

If you're interested in learning more about work, you can check out our code and data here!