Exploratory Data Analysis and Visualization for Art Auction Data

Alexandra Sudomoeva, Elizabet Doliar, Serena Zhang, Basil Vetas

4/26/2018

Introduction

Introduction

For our project we chose to perform an exploratory analysis and visualization on a dataset containing the historical auction prices of thousands of pieces of art. We chose this topic for a variety of reasons—whether you come from the arts and humanities, or economics and finance, art is an interesting and complex field. With new data starting to become available, for the first time we have the ability to analyze and better understand this historically opaque industry. It is a potential treasure trove for data scientists.

For hundreds of years fine art has been valued by cultures throughout the world for its aesthetic and beauty, but fine art as an asset class also appeals to many people today as an attractive alternative investment that is largely non-correlated to the traditional stock market. The fine art asset class is estimated to have a global market value of over $60 billion in annual turnover and has historically outperformed the S&P 500 in terms of compound annual growth (specifically, 15-year CAGR). Geographically, the art market has for decades been dominated by the United States, United Kingdom and China with the greatest concentration of art sales in the entire world occurring right here in New York City.

Each of our team members has a diverse background. In addition to experience as data scientists, a few of us have a background in economics and have worked for financial services and tech companies. A few of us have also worked in the art industry for galleries or collectors. This topic was a cross-section of our experiences and interests that allowed us to explore timely questions about the prices and dynamics of the fine art market.

From a high level, the questions that intrigued us most were related to knowing more about the auction sale prices (called hammer prices) for pieces of art (called lots) in terms of different artists, location and time periods. We initially developed a large list of potential questions, and narrowed these down based on viability in terms of the data we had available as well as the scope and time constraints for the project. Some of the specific questions we decided to look at are:

1. What is the distribution of quantity of lots and auctions by location, year, and season?
2. What are the lot titles that have higher price or appear more?
3. How is price related to artist's era/when he/she was born?
4. Do auction prices vary by location and season?
5. Did Financial Crisis have any effect on the auctions?
6. Does the order in which the lot is presented affect the overall sell price?

Team Members & Contributions

Each team member worked equally hard to bring this project to life. While having weekly progress meetings and brainstorming sessions, we have adopted a divide-and-conquer approach around building graphs and the corresponding write-up. After writing each of their respective sections, all team members contributed to editing and revising the final report.

The Introduction as well as Data Description sections were written by *Basil Vetas*. Basil acquired the dataset used in the project, and contributed to data preprocessing. In addition, Basil created the interactive component of the assignment.

Data Quality Analysis as well as some of the data preprocessing was performed by *Alexandra Sudomoeva*. Alexandra also built visualizations for the Financial Crisis' impact in the main analysis. Lastly, she assisted in writing the instructions for the interactive component.

Serena Zhang did the majority of the data preprocessing, and also worked on the main analysis as well as the Executive Summary. Her analysis included auction price correlation with lot titles, artist's era, order in auction, season and location.

The main contributor to the analysis around basic information was *Elizabet Doliar*. She also worked at analyzing project titles and creating an auction price summary by building a parallel coordinates plot. Lastly, Elizabet helped put together the executive summary.

Data Description

Overview

This dataset was provided by Arthena (https://arthena.com/ (https://arthena.com/)), a startup based in New York that uses data to analyze and invest in art. Arthena originally sourced the data from Sotheby's historical auction data via the scraping of public web pages. Our raw dataset

includes 22711 rows of data from Sotheby's auctions. Each row represents an individual lot from an auction (a lot could be an individual painting, a sculpture, or sometimes even a collection of works). The raw dataset includes 25 columns of data. Each column represents a feature related to either that specific lot, the artist of the lot, or the auction where the lot was sold. Column definitions are listed below by category (Lot, Auction or Artist). For our analysis we also derive new columns from the original raw dataset. The derived column definitions are also listed below. Including these columns we ended with a total of 43 columns of data. The detailed logic and thought process is described more in the 'Data Quality Analysis' section.

```
#Read data
art_df = read.csv("final_sothebys.csv", header=TRUE, na.strings=c("", NA))
```

Column Definitions:

```
#Dropping features with no interest (non-informative) in the analysis
drop <- c("X", "Unnamed..0", "provenance", "auc_desc", "auction_house_id", "external_image_url", "literature", "e
nd_date")
art_df <- art_df[, !(names(art_df) %in% drop)]
```

Lot
lot_id: a unique id for each lot.
lot_title: the title of the lot. A lot can sometimes consist of multiple pieces of art. We assume that 1 piece is 1 lot since that is most common.
estimate_low: the low-end auction price estimate for a lot, given by Sothebys.
estimate_high: the high-end auction price estimate for a lot, given by Sothebys.

hammer_price_bp: how much the lot was sold for at auction, plus buyers premium (a percentage fee taken by Sothebys and paid by the buyer).
currency: currency denomination for the price estimates and hammer price (limited to USD, EUR, GBP, HKD).
nth_in_auction: the order that the lot was presented in at auction.
lot_number: a number assigned to a lot for the given auction, different than nth_in_auction.
condition: description of the condition of the lot (messy text field - not used for our analysis).
provenance: description of who owned the lot previously (messy text field - not used for our analysis).
literature: different publications that the lot was mentioned in (messy text field - not used for our analysis).
external_image_url: link to the image (not used for our analysis).

Auction
auction_house_id: unique id for each auction house (for this dataset, all 1 since we are only using Sothebys data).
auction_id: unique id for each auction.
auc_title: title of the auction.
number_of_lots: total number of lots in the auction.
location: location where the auction was held.
start_date: start date of the auction.
end_date: end date of the auction (same as start date for this dataset).
auc_desc: description of the auction (messy text field - not used for our analysis).
sale_id: unique auction sale id assigned by Sothebys.

Artist
artist_id: unique id for each artist.
name: name of the artist.
birth_year: artist's approximate birth year (messy text field - not used for our analysis).
death_year: artist's approximate death year (messy text field - not used for our analysis).

Derived
estimate_avg: the average between estimate_low and estimate_high.
is_untitled: an indicator variable whether the name of the lot is "untitled" (in some language).
auc_year: the year of the auction (YYYY format).
auc_month: the month of the auction (as integers 1-12).
auc_season: the season of the auction (as integers 1-4).
auc_date: the date of the auction.
auth_era: groupe the birth year of authors into 10-year perios
nth_in_auction: the order in the auction by quantiles (as integers 1-100).
percent_in_auction: the percentage through an auction that a lot was shown (nth_in_auction divided by number_of_lots).
hammer_price_bp_usd: hammer_price_bp converted to usd.
estimate_high_usd: estimate_high converted to usd.

estimate_low_usd: estimate_low converted to usd.
estimate_avg_usd: estimate_avg converted to usd.
hammer_price_bp_usd_range: factor of bucketed hammer price ranges in usd.

Data Quality Analysis

Preprocessing

In this part of the project, we will be exploring the quality of the data provided for the auction data. Before analyzing the overall quality, a simple preprocessing was conducted in Python that included the following steps:

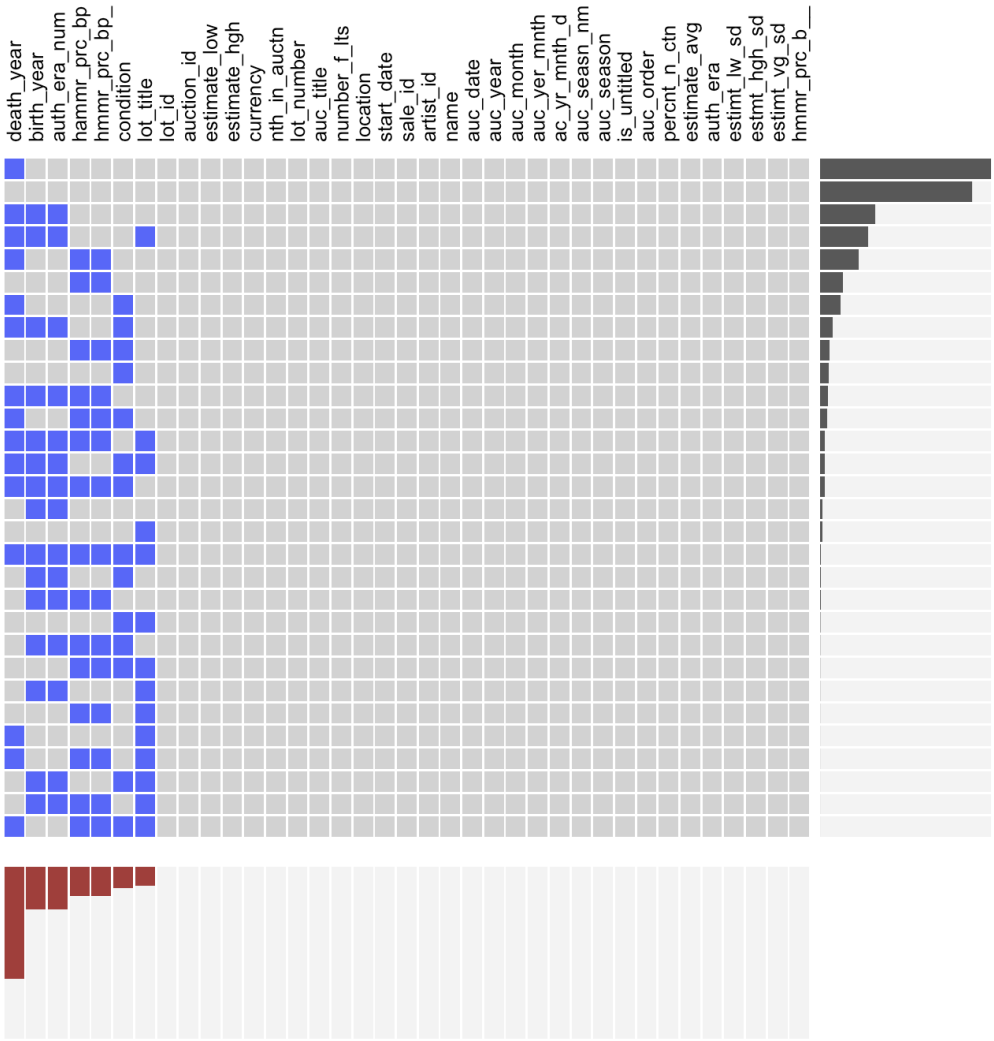
1. Since ‘Start Date’ and ‘End Date’ are always the same within the dataset, we decided to only use ‘Start Date’ for the time when the auction occurred. Converted the column to DateTime type.
2. Based on ‘Start Date’ column, we added columns for normalized year, month, and season.
3. The “Title” column accepted a variety of title across different languages. We added a new column to indicate whether the piece is “untitled” while checking for the top 5 most common languages used: Italian, Dutch, English, French, Spanish, German.
4. Created a new column ‘nth_in_auction’ that would break the order of the auction into 100 tiles and a column ‘percent_in_auction’ to show the percentage of the order within a specific auction.
5. Added ‘Average Estimate’ column to reflect the average price estimation between the high and the low
6. Added column ‘auth_era’ to group author’s birth year for every 10 years.
7. Converted all currency to USD to be able to compare pieces sold across different locations. We matched the exchange rate at the time of the sell to properly convert all transactions.
8. Added ‘hammer_price_bp_usd_range’ that allocates the hammer prices into 4 ranges.

The complete preprocessing python notebook file can be found here: https://github.com/serenazzz/art-auction-visualization-project/blob/master/Preprocessing.ipynb (https://github.com/serenazzz/art-auction-visualization-project/blob/master/Preprocessing.ipynb)

Data Quality Exploration

We began by looking at all the missing values and if there are any general patterns. Since we were dealing with a large amount of rows, we reduced repeated patterns to one row using the visna function.

```
#install.packages("extracat")
library(extracat)
visna(art_df, sort= "b")
```



Looking at the output above, we can conclude that the overall state of the dataset is relatively good. The second most common pattern has no missing values while the two most “problematic” features appear to be death_year and birth_year.

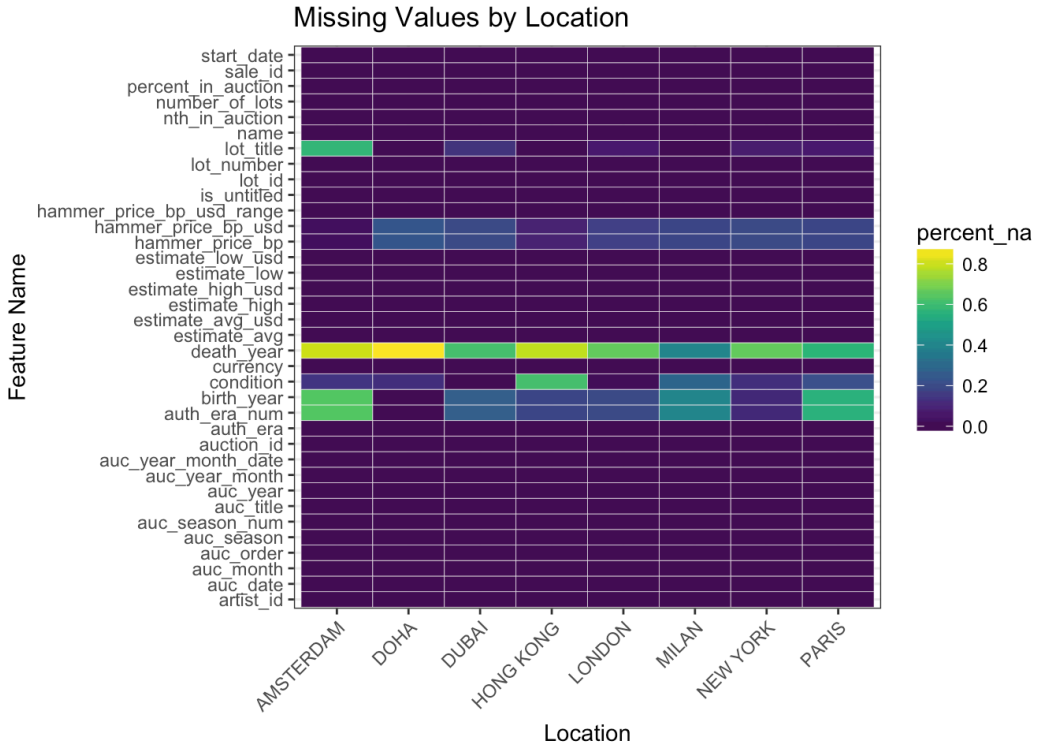
One important observation we were careful about during further analysis is the fact that a relatively significant amount of missing values under hammer_price_bp feature. We have explored qualitative reasons behind the missing values in that category with the provider of the data. After careful observation, we found out that the missing values are actually caused by two auction locations that might have less strict regulations around data governance (Doha and Dubai). The high number of NAs in New York is left unidentified. This observation is outlined in the table below.

```
art_price <- art_df[, c("location", "hammer_price_bp")]
percent_missing <- art_price %>% group_by(location) %>%
  summarise(num_na = sum(is.na(hammer_price_bp)), total = n()) %>%
  mutate(percent_na = num_na/total)%>%
  arrange(-percent_na)
percent_missing
```

```
## # A tibble: 8 x 4
##   location  num_na total percent_na
##   <fct>      <int> <int>      <dbl>
## 1 DOHA          16    67     0.239
## 2 NEW YORK    1988  9663     0.206
## 3 DUBAI         11    54     0.204
## 4 MILAN        253  1381     0.183
## 5 PARIS        465  2542     0.183
## 6 LONDON      1058  6311     0.168
## 7 HONG KONG     37   456     0.0811
## 8 AMSTERDAM     70  2216     0.0316
```

We also looked at the percentage of missing value by location for all attributes to check if other features were impacted.

```
#install.packages("viridis")
library(viridis)
art_location <- art_df %>% gather(attribute, value, -location)
percent_missing <- art_location %>% group_by(location, attribute) %>%
  summarise(num_na = sum(is.na(value)), total = n()) %>%
  mutate(percent_na = num_na/total)
ggplot(percent_missing, aes(x = location, y = attribute, fill = percent_na)) +
  geom_tile(color = "white") +
  ggtitle("Missing Values by Location") +
  xlab("Location") + ylab("Feature Name") +
  scale_fill_viridis() +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This map allows us to see that the missing values for some features like hammer_price_bp, condition, and birth_year are actually only missing at a high rate in certain locations.

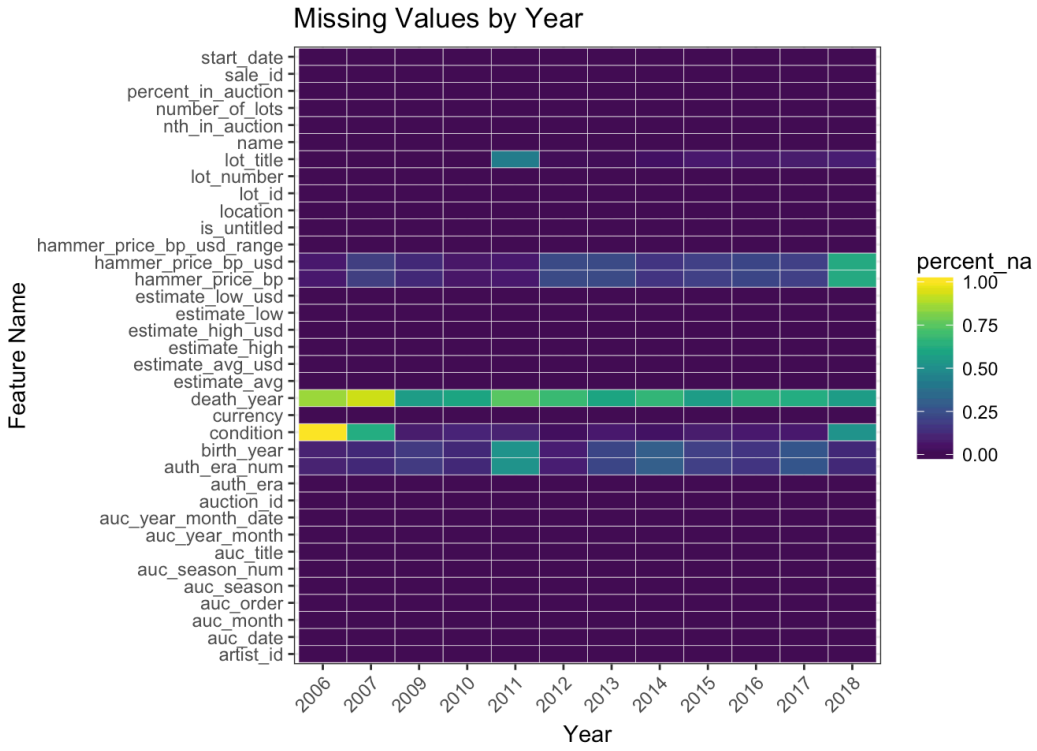
Indeed, when looking at the table below, the overall percentage of missing values across different locations varies quite significantly. Amsterdam is the top location with most missing values (8% of total). Therefore, we can speculate that this data could be MAR (missing at random) depending on a location feature.

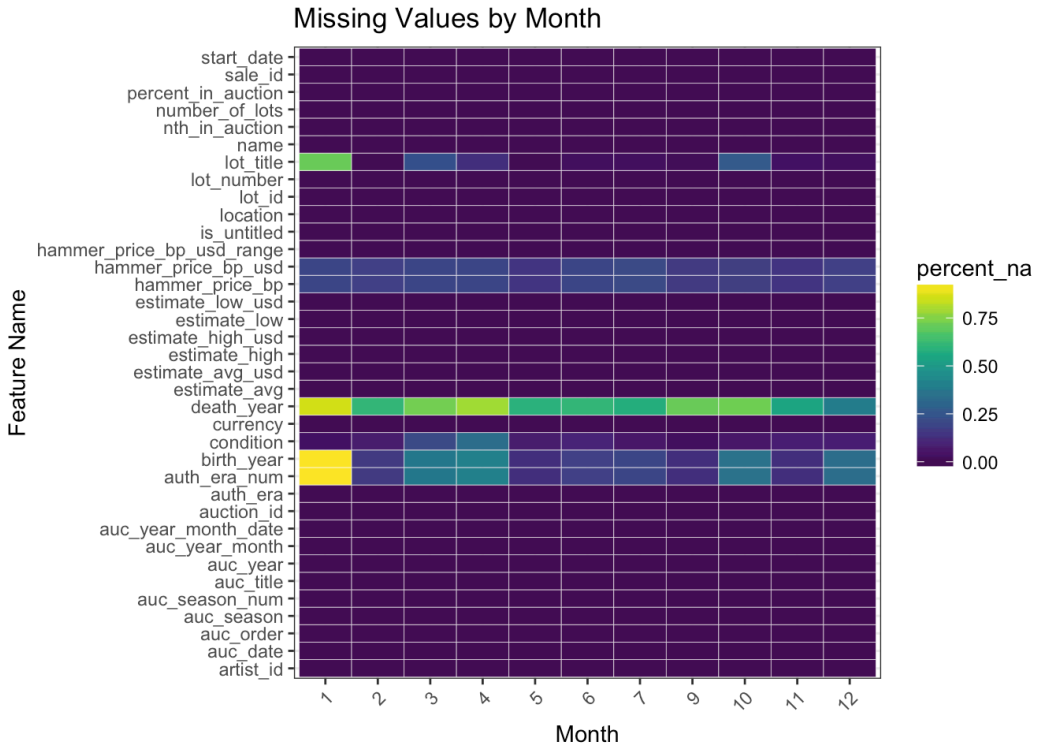
```
art_location <- art_df %>% gather(attribute, value, -location)
percent_missing <- art_location %>% group_by(location) %>%
  summarise(num_na = sum(is.na(value)), total = n()) %>%
  mutate(percent_na = num_na/total)%>%
  arrange(-percent_na)
percent_missing
```

```
## # A tibble: 8 x 4
##   location  num_na  total percent_na
##   <fct>      <int>  <int>      <dbl>
## 1 AMSTERDAM   6308  79776     0.0791
## 2 PARIS       5873  91512     0.0642
## 3 HONG KONG    890  16416     0.0542
## 4 MILAN       2547  49716     0.0512
## 5 DUBAI         90   1944     0.0463
## 6 LONDON      9279 227196     0.0408
## 7 NEW YORK   14146 347868     0.0407
## 8 DOHA          98   2412     0.0406
```

This leads to a logical question. Are there any other variables that could explain the missing data? Therefore, we also looked at similar graphs while grouping by year and month.

```
art_year <- art_df %>% gather(attribute, value, -auc_year)
year_missing <- art_year %>% group_by(auc_year, attribute) %>%
  summarise(num_na = sum(is.na(value)), total = n()) %>%
  mutate(percent_na = num_na/total)
year <- ggplot(year_missing, aes(x = factor(auc_year), y = attribute, fill =
                                   percent_na)) +
  geom_tile(color = "white") +
  ggtitle("Missing Values by Year") +
  xlab("Year") + ylab("Feature Name") +
  scale_fill_viridis() +
  theme_bw()+theme(axis.text.x = element_text(angle = 45, hjust = 1))

art_month <- art_df %>% gather(attribute, value, -auc_month)
month_missing <- art_month %>% group_by(auc_month, attribute) %>%
  summarise(num_na = sum(is.na(value)), total = n()) %>%
  mutate(percent_na = num_na/total)
month <-ggplot(month_missing, aes(x = factor(auc_month), y = attribute, fill =
                                  percent_na)) +
  geom_tile(color = "white") +
  ggtitle("Missing Values by Month") +
  xlab("Month") + ylab("Feature Name") +
  scale_fill_viridis() +
  theme_bw()+theme(axis.text.x = element_text(angle = 45, hjust = 1))
grid.arrange(year, month)
```





It comes as no surprise that older years (2006-2009) carry a lot of the missing values for features around description and condition (there is no data for 2008). There also looks to be certain periods with relatively clean data: 2012-2013 and 2015-2017. Another interesting observation is the fact that certain seasons (Winter and beginning of Spring) tend to have more missing values than other months (there is no data for July).

Next, let us look at the exact percentages and values for the overall missing data.

```
#install.packages("skimr")
library(skimr)
skimr::skim(art_df) %>% filter(stat =="missing") %>% arrange(desc(value)) %>% select(variable, value) %>% mutate(
percent = value/nrow(art_df)) %>% filter (percent>0)
```

```
## # A tibble: 7 x 3
##   variable            value percent
##   <chr>               <dbl>   <dbl>
## 1 death_year          14806   0.653
## 2 birth_year           5676   0.250
## 3 auth_era_num         5676   0.250
## 4 hammer_price_bp      3898   0.172
## 5 hammer_price_bp_usd  3898   0.172
## 6 condition            2808   0.124
## 7 lot_title            2469   0.109
```

Looking at the table, death and birth year are missing more than 25% of their data. More importantly, hammer_price is missing nearly 20%. Based on what we saw when looking by location, our guess is that some auction ids are missing the hammer_price_bp in its entirety and hence the difference. It can be easily checked by looking at the aggregate table.

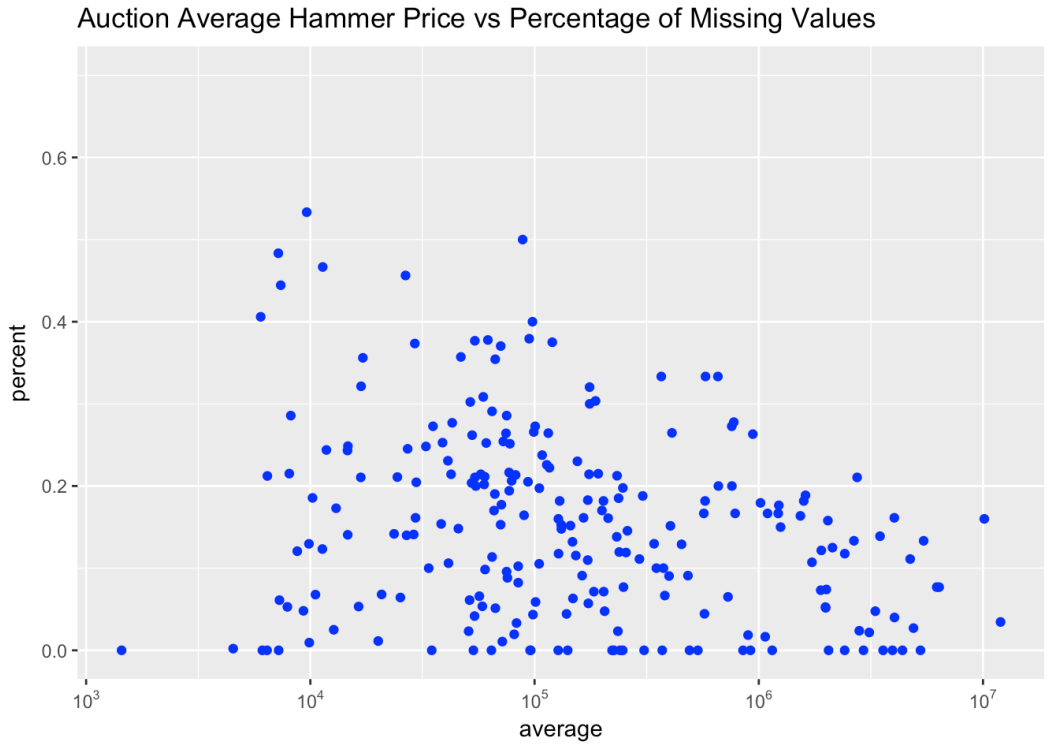
```
art_price <- art_df[, c("auction_id", "hammer_price_bp")]
percent_missing <- art_price %>% group_by(auction_id) %>%
  summarise(num_na = sum(is.na(hammer_price_bp)), total = n()) %>%
  mutate(percent_na = num_na/total)%>%
  arrange(-percent_na)
percent_missing %>% filter(percent_na>=0.4)
```

```
## # A tibble: 10 x 4
##    auction_id num_na total percent_na
##         <int>  <int> <int>      <dbl>
##  1          2    258   258      1.00
##  2          3    173   173      1.00
##  3        108     48    90      0.533
##  4        180      8    16      0.500
##  5        100    116   240      0.483
##  6        186      7    15      0.467
##  7        133     47   103      0.456
##  8         30     36    81      0.444
##  9         20    136   335      0.406
## 10         19     14    35      0.400
```

Looking at the data table, there is a significant number of auctions that are missing more than 40% of the price data (sometimes even 100%). Therefore, it must be that not only the locations but also the auction itself is a determining factor in missing hammer price value.

Lastly, we checked for a correlation between average hammer_price_bp and percentage of missing values (since it is the one feature that matters the most in our analysis and is most susceptible to such correlations). Could it be that very high/low lots are simply not reported and, therefore, are missing more?

```
art_auction <- art_df[, c("auction_id", "hammer_price_bp")] %>% arrange(auction_id)
art_average <-art_auction %>% filter(!is.na(hammer_price_bp)) %>%
  group_by(auction_id) %>% summarise(mean =mean(hammer_price_bp)) %>% arrange(auction_id)
percent_missing <- percent_missing %>% arrange(auction_id)
add_2 <- data.frame(auction_id=2, mean=0)
add_3 <- data.frame(auction_id=3, mean=0)
art_average <- rbind(art_average, add_2)
art_average <- rbind(art_average, add_3) %>% arrange(auction_id)
percent <- percent_missing[4]
average <- art_average[2]
auction_id <- percent_missing[1]
corr <- data.frame(auction_id=auction_id, percent=percent, average = average)
ggplot(corr, aes(average, percent)) + geom_point(col="blue") +ggtitle("Auction Average Hammer Price vs Percentage
of Missing Values") +
  scale_x_log10( breaks = scales::trans_breaks("log10", function(x) 10^x),
   labels = scales::trans_format("log10", scales::math_format(10^.x)))+ylim(0,0.7)#cutting the y limit since only
one point has >70% missing
```



There does not seem to be very strong correlation between the two variables. Just a subtle suggestion that auctions with smaller average price tend to have more NAs. Therefore, we only consider location as the main determinant for missing values around hammer_price_bp.

Before we move on, let us quickly summarize the findings from the data quality analysis:

1. filter out NAs for price estimates
2. take a note to exclude Doha and Dubai fro hammer_price_bp analysis due to low base high NAs

```
# construct final dataset based on the quality analysis
art_final <- art_df %>% filter(!is.na(estimate_low)) %>%
            drop_na(hammer_price_bp_usd) %>%
          filter(location %in%  c("HONG KONG", "NEW YORK", "LONDON", "PARIS", "MILAN","AMSTERDAM"))
```

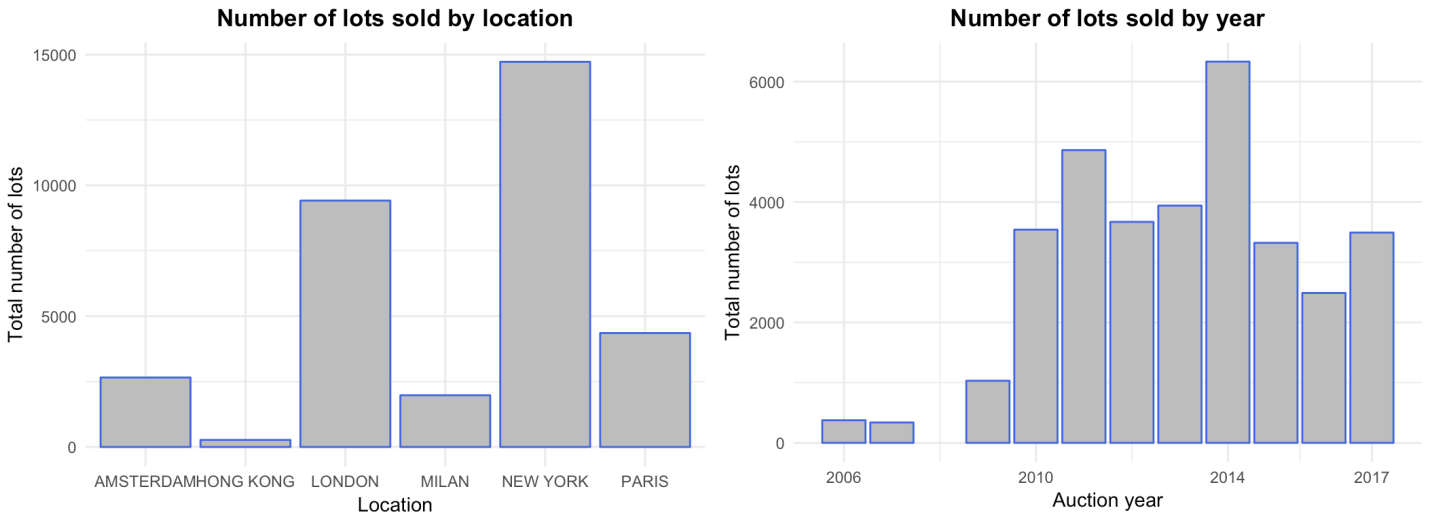
Main Analysis (Exploratory Data Analysis)

Our research can be divided into three sections: General questions about the auction information, understanding fluctuations in lot prices and its correlation with different variables

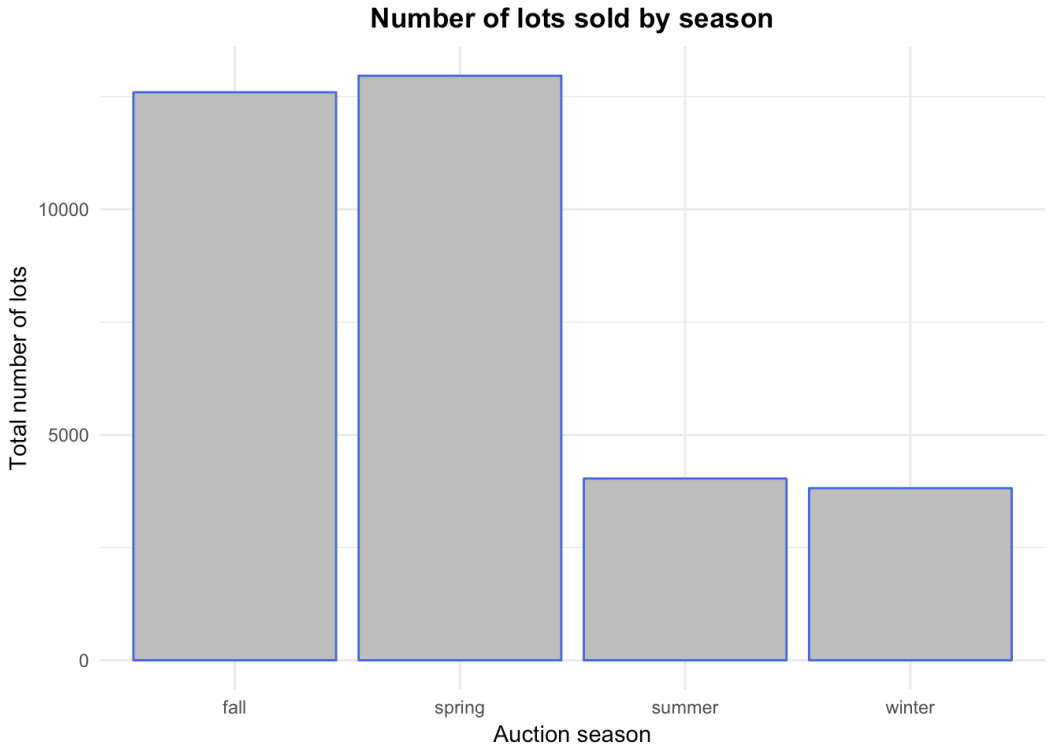
General Auction Information

We started by asking many questions about possible relationships between variables. The first set of plots will explore the correlation between number of lots sold and year, location and season. We were hoping to notice meaningful trends that can be further explored in subsequent sections. Since we had only a few locations, seasons and years we chose a bar chart and excluded duplicate rows by the ‘number of lots’ column.

```
library(gridExtra)
##Lots by location
art_info <- subset(art_final, select=c( "location", "number_of_lots" ))
art_info <- art_info[!duplicated(art_info$number_of_lots),]
art_group <- art_info %>% group_by(location)%>% summarise(B=sum(number_of_lots))
p1 <- ggplot(art_group, aes(x= location, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")  +labs(x = "Location")+labs(y = "Total number of l
ots") + ggtitle("Number of lots sold by location")  + theme_minimal() + theme(plot.title = element_text(hjust = 0
.5,face="bold"))

art_info2 <- subset(art_final, select=c( "auc_year", "number_of_lots" ))
art_info2 <- art_info2[!duplicated(art_info2$number_of_lots),]
art_info2 <- art_info2 %>% group_by(auc_year)%>% summarise(B=sum(number_of_lots))
p2 <- ggplot(art_info2, aes(x= auc_year, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")  +labs(x = "Auction year")+labs(y = "Total number
of lots") + ggtitle("Number of lots sold by year")  + theme_minimal()+ theme(plot.title = element_text(hjust = 0.
5,face="bold"))+ scale_x_continuous(breaks= c(2006, 2010, 2014, 2017))

grid.arrange(p1, p2, nrow = 1)
```



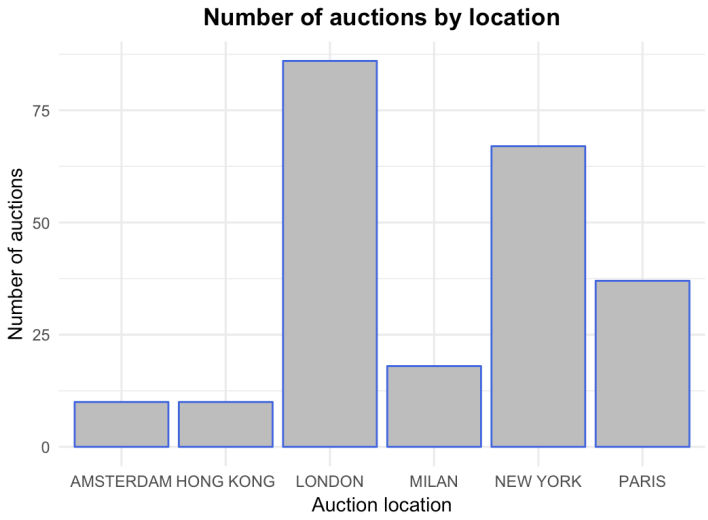

```
##Lots by season
art_info <- subset(art_final, select=c( "auc_season", "number_of_lots" ))
art_info <- art_info[!duplicated(art_info$number_of_lots),]
art_group <- art_info %>% group_by(auc_season)%>% summarise(B=sum(number_of_lots))
ggplot(art_group, aes(x= auc_season, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")  +labs(x = "Auction season")+labs(y = "Total numbe
r of lots") + ggtitle("Number of lots sold by season") + theme_minimal() + theme(plot.title = element_text(hjust
= 0.5,face="bold"))
```

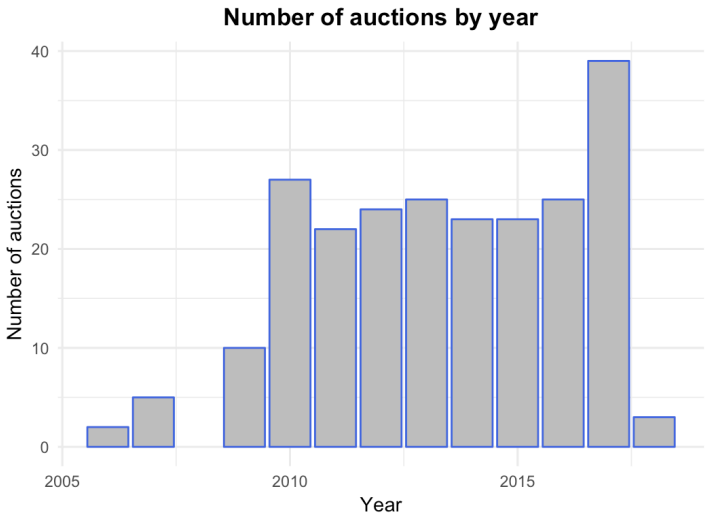


The graphs above suggest that fall and spring are the most popular seasons to acquire a masterpiece. At the same time, there was a significant increase in number of lots sold since 2010, suggesting that investors started to see the art market as a form of long-term non-liquid investment after the financial crisis of 2008 (we will explore this hypothesis in more detail later in the analysis). The main hubs for art exchanges formed in London, New York and Paris.

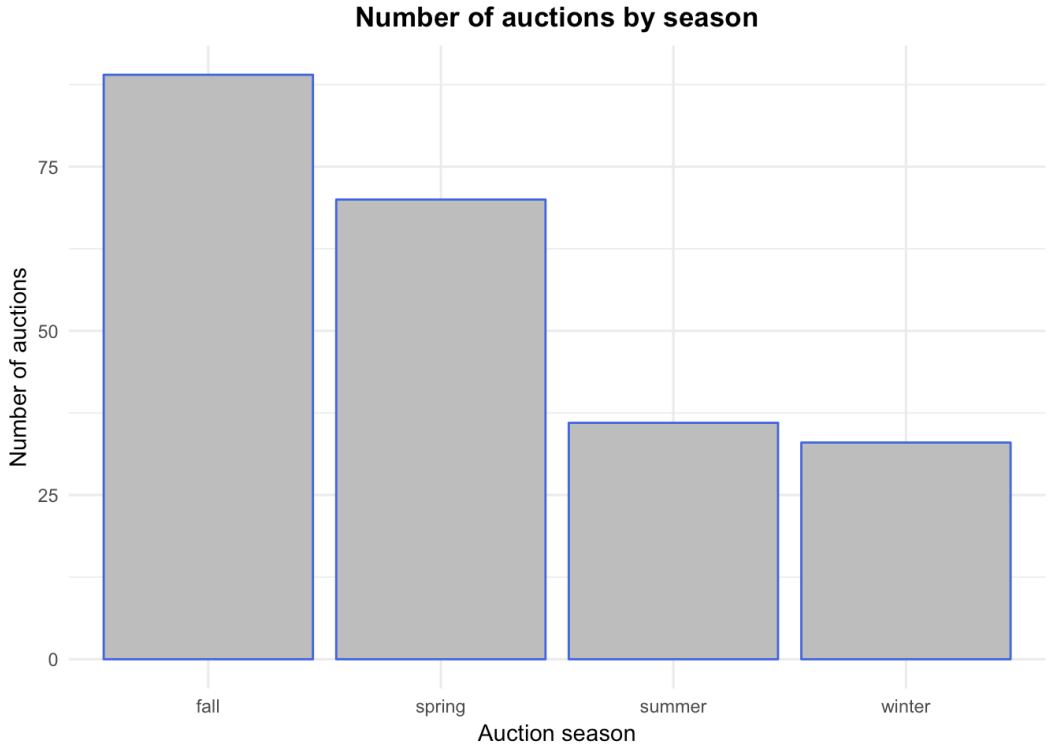
The next set of graphs will focus on the number of actions by year, location and season.

```
##Auctions by location, year, season
art_info <- subset(art_final, select=c( "auction_id", "location" ))
art_info <- art_info[!duplicated(art_info$auction_id),]
art_group <- art_info %>% group_by(location)%>% count(auction_id) %>% summarise(B=sum(n))
p1<-ggplot(art_group, aes(x= location, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")  +labs(x = "Auction location")+labs(y = "Number of
auctions") + ggtitle("Number of auctions by location") + theme_minimal()+ theme(plot.title = element_text(hjust =
0.5,face="bold"))

art_info2 <- subset(art_final, select=c( "auction_id", "auc_year" ))
art_info2 <- art_info2[!duplicated(art_info2$auction_id),]
art_info2 <- art_info2 %>% group_by(auc_year)%>% count(auction_id) %>% summarise(B=sum(n))

p2<- ggplot(art_info2, aes(x= auc_year, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")+labs(y = "Number of auctions")+labs(x = "Year") +
theme_minimal() + ggtitle("Number of auctions by year") + theme(plot.title = element_text(hjust = 0.5,face="bold"
))
grid.arrange(p1, p2, nrow = 1)
```





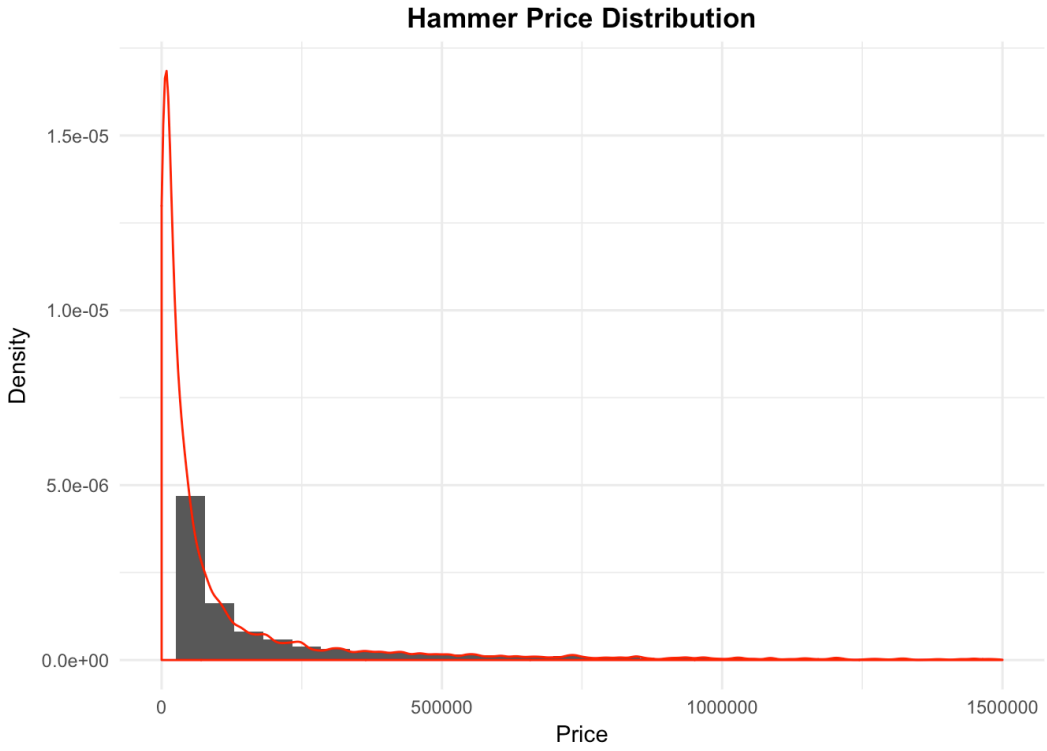

```
##Auctions by season
art_info <- subset(art_final, select=c( "auction_id", "auc_season" ))
art_info <- art_info[!duplicated(art_info$auction_id),]
art_group <- art_info %>% group_by(auc_season)%>% count(auction_id) %>% summarise(B=sum(n))
ggplot(art_group, aes(x= auc_season, y = B)) +
    geom_bar( stat='identity', color="royalblue", fill="grey")  +labs(x = "Auction season")+labs(y = "Number of a
uctions") + ggtitle("Number of auctions by season") + theme_minimal() + theme(plot.title = element_text(hjust = 0
.5,face="bold"))
```



Plotting the number of auctions based on location, season and year turned out to have similar results to the previous section and confirmed ideas suggested above. Namely, fall and spring are the two main auction seasons. London, New York and Paris are the main locations for art trading while New York has more lots sold and London has more auctions overall. And after 2009 the number of auctions conducted yearly around the world increased significantly.

To answer our second set of questions and evaluate the different triggers of variability in art prices we decided to explore fluctuations in Hammer Prices. Therefore, we created a histogram with density curve to visualize the distribution of Hammer Price.

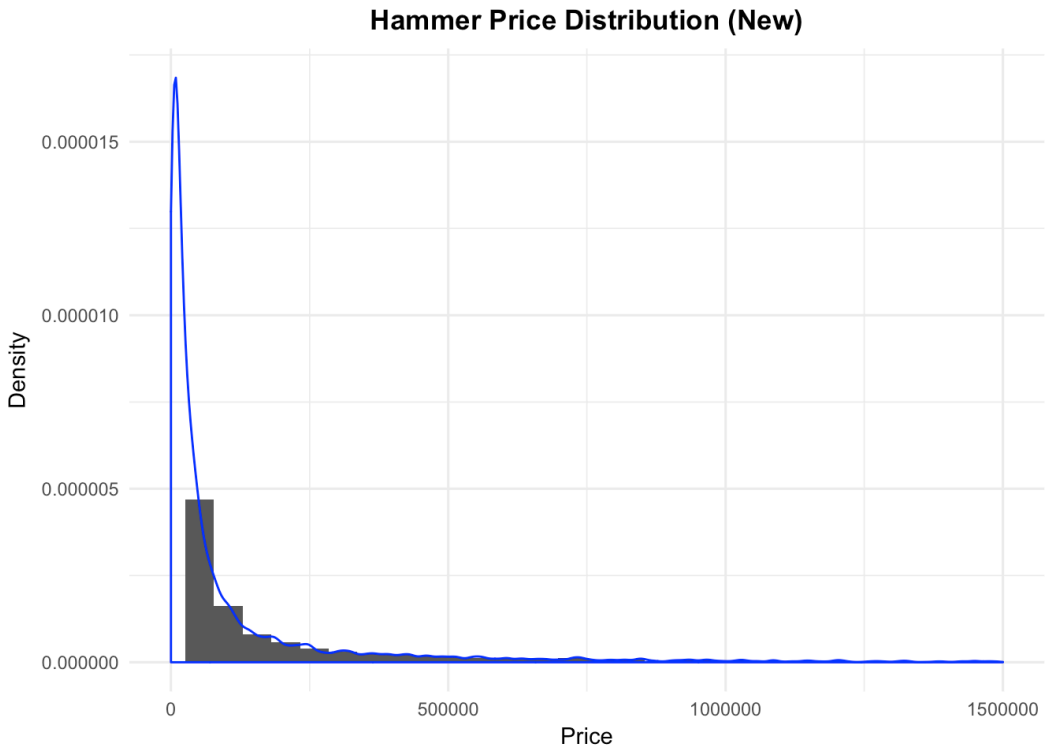
```
##Hammer Price
ggplot(art_final, aes(x= hammer_price_bp_usd)) + geom_histogram(aes(y=..density..))  + geom_density(col="red") +
xlim(0,1500000) + theme_minimal() +xlab("Price") +ylab("Density") +ggtitle("Hammer Price Distribution")+ theme(pl
ot.title = element_text(hjust = 0.5,face="bold"))
```



Distribution of Hammer Price is skewed to the right and has a long tail. We believe the reason for this is various price ranges for different groups of lots sold on the market. Variability in price can be explained by difference in art genres (Contemporary vs Renaissance for example), quality, age and popularity of the artwork. These factors can distinguish a masterpiece sold for $135 million like “Portrait of Adele Bloch-Bauer” by Gustav Klimt and Untitled painting by Mark Rothko sold for only 6.5 million.

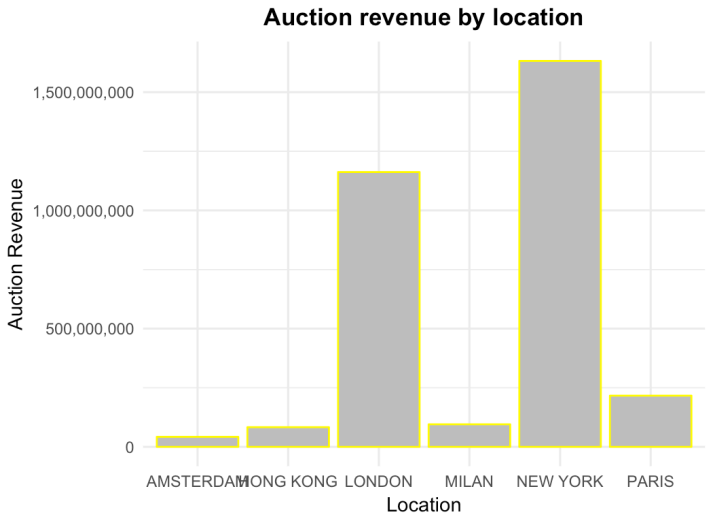
Constructing valuable models in the next parts of our research fully depend on the ability to manipulate hammer price in the right way. We decided to remove the outliers.

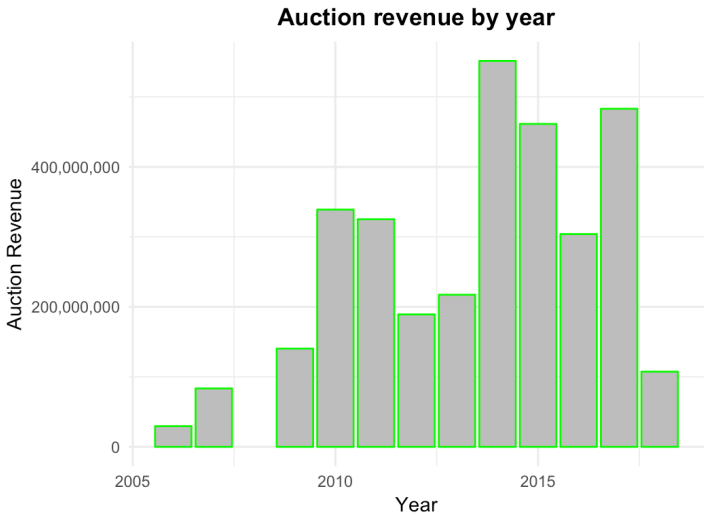
```
art_final <-art_final%>%
          filter(abs(art_final$hammer_price_bp_usd -
               median(art_final$hammer_price_bp_usd)) <=3*sd(art_final$hammer_price_bp_usd))

ggplot(art_final, aes(x= hammer_price_bp_usd)) + geom_histogram(aes(y=..density..))  + geom_density(col="blue") +
xlim(0,1500000)+ theme_minimal() +xlab("Price") +ylab("Density") +ggtitle("Hammer Price Distribution (New)")+ the
me(plot.title = element_text(hjust = 0.5,face="bold"))+scale_y_continuous(labels = scales::comma)
```



Adjusted hammer price brought our attention to distribution of revenue over time and location. For both plots we chose a bar chart.

```
##revenue by location

MyData <- subset(art_final, select=c( "location", "auction_id", "hammer_price_bp_usd" ))


MyData5 <- MyData %>% group_by(location)%>% summarise(B=sum(hammer_price_bp_usd))


p1 <- ggplot(MyData5, aes(x= location, y = B)) +
    geom_bar( stat='identity', color="yellow", fill="grey")+labs(y = "Auction Revenue")+labs(x = "Location") + gg
title("Auction revenue by location") + theme_minimal() + theme(plot.title = element_text(hjust = 0.5,face="bold")
)+scale_y_continuous(labels = scales::comma)

##revenue by year

MyData_1 <- subset(art_final, select=c( "auc_year", "auction_id", "hammer_price_bp_usd" ))

MyData_1 <- MyData_1 %>% group_by(auc_year)%>% summarise(B=sum(hammer_price_bp_usd))

p2<-ggplot(MyData_1, aes(x= auc_year, y = B)) +
    geom_bar(stat='identity', color="green", fill="grey")+labs(y = "Auction Revenue")+labs(x = "Year") + ggtitle(
"Auction revenue by year") + theme_minimal()  + theme(plot.title = element_text(hjust = 0.5,face="bold"))+scale_y
_continuous(labels = scales::comma)

grid.arrange(p1, p2, nrow = 1)
```



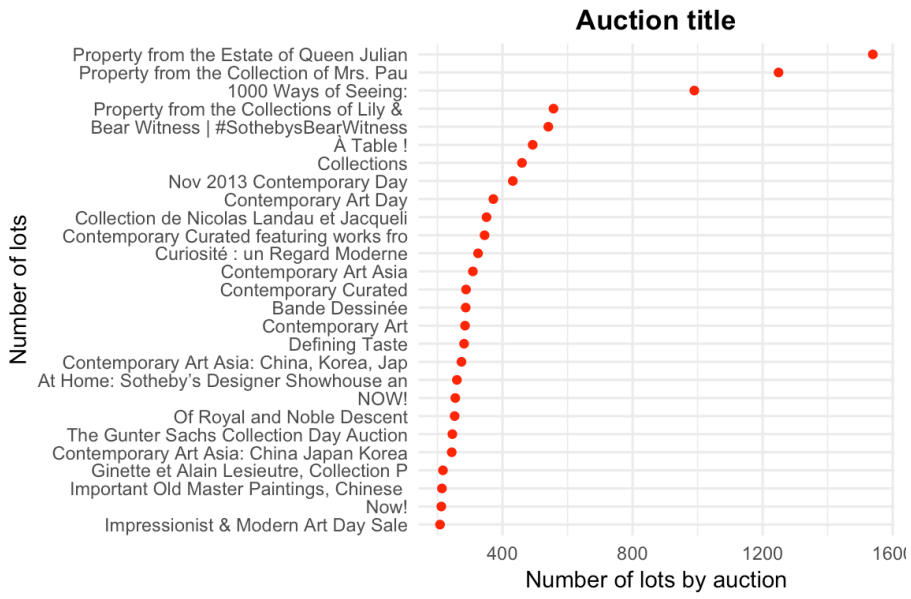


After the financial crisis world art revenues went up following the assumption that people saw art as a form of investment. In 2013 revenues declined again possibly due to a slow down in the art market and went back up in the following years. London and New York continue to lead the way as the main centers for the exchange of art.

Another question that we thought would help us explore the data: Variability of lots across auctions?

In the data we had 89 auctions varying in theme, concept and length. We were particularly interested in the genres of auctions that had the largest number of lots. We created a Cleveland dot plot and filtered by auctions that had more than 200 lots.

```
##Cleveland dot plot
MyData2 <-  strtrim(art_final$auc_title, 40)
art_final$auc_title <- MyData2

MyData1 <- art_final[!duplicated(art_final$auc_title),]

MyData3 <- subset(MyData1, select=c("auc_title", "number_of_lots"))

MyData3$auc_title <- factor(MyData3$auc_title, levels = MyData3$auc_title[order(MyData3$number_of_lots)])

MyData3<-  MyData3[which(MyData3$number_of_lots>200),]

ggplot(MyData3, aes( x = auc_title, y = number_of_lots)) + geom_point(stat="identity",  color="red") + coord_flip
()+ theme_minimal() +labs(y = "Number of lots by auction")+labs(x = "Number of lots") + ggtitle("Auction title")
+ theme(plot.title = element_text(hjust = 0.5,face="bold"))
```

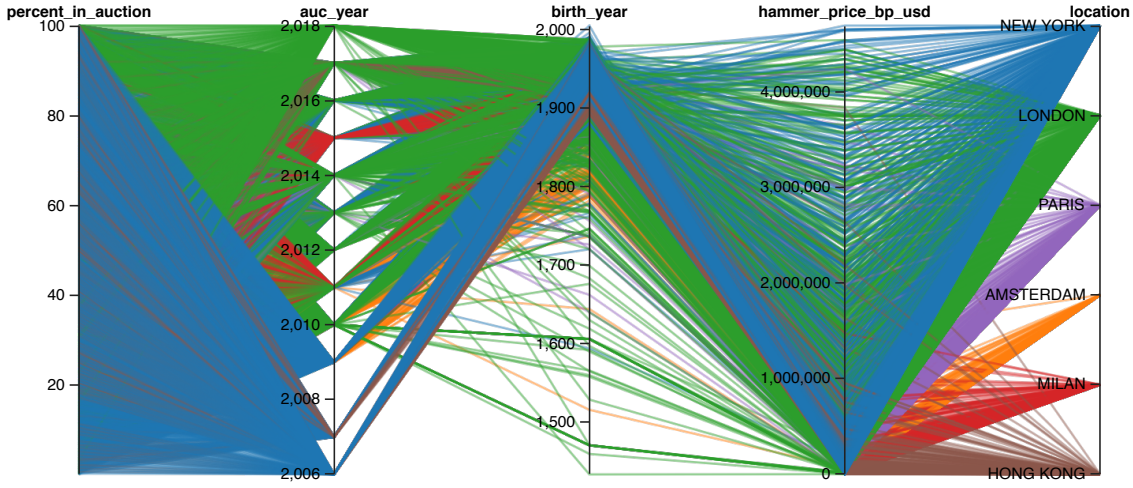


Based on the graph we identified the theme of each auction and calculated shares of different themes. The result showed that 48% of the auctions with the highest number of lots are Contemporary auctions. This results can be easily explained since modern artists create more and more content every year while for older masters artworks are main resold (nothing new is generated).

To conclude exploration of data we created a parallel coordinate plot for auction year, birth year of the artist, lot number in the auction, and hammer price colored by location.

```
MyDatas <-subset(art_final, select=c( "percent_in_auction", "auc_year" , "birth_year", "hammer_price_bp_usd", "lo
cation"))
MyDatas<- MyDatas %>%
  filter(birth_year>1400)
```

```
#devtools::install_github("timelyportfolio/parcoords")
library(parcoords)
#library(httpuv)

parcoords(MyDatas,
    rownames = F
    , brushMode = "1D-axes", alpha =0.5,color = list(
      colorBy = "location", colorScale = htmlwidgets::JS("d3.scale.category10()"))
    )
```



From this chart, we can see that most of New York's lots have author born after 1900. However, other than that, we can't read much valuable inforamtion in this graph as the correlations are not strong enough and lines are overlapping.

Do Certain Lot Attributes Result in Higher Price?

Lot Titles

What lots have higher price?

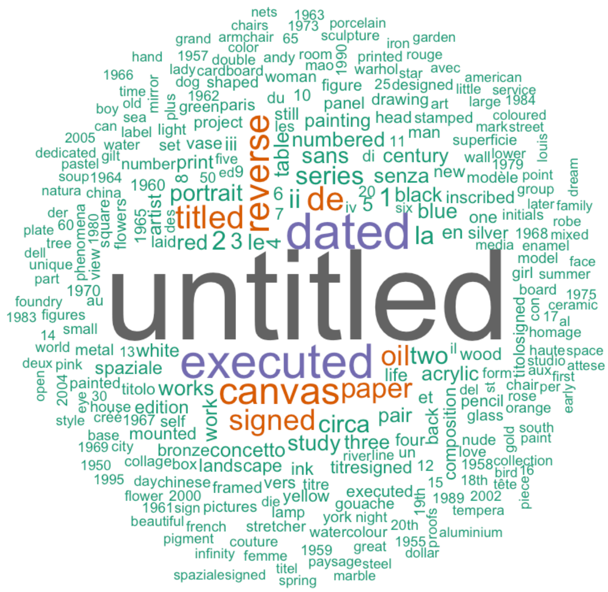
To begin the analysis, we want to see the titles of expensive lots and plot their size with respect to the prices.

```
df1 <- art_final
df_wordcloud <- df1[,c("lot_title","hammer_price_bp_usd")]
df_wordcloud <- arrange(df_wordcloud,desc(df_wordcloud$hammer_price_bp_usd))[1:1000,]
library(wordcloud)
library(tm)
pal <- brewer.pal(9, "OrRd")
pal <- pal[-(1:3)]
wordcloud(df_wordcloud$lot_title, min.freq = 10,df_wordcloud$hammer_price_bp_usd, scale=c(5, .5), random.order =
FALSE, random.color = FALSE, colors= pal)
```

A lot of expensive lots appear to have “untitled” in their title. To get a better idea of what words artists tend to name their pieces, we count the frequency of each word appear in all titles and see which words do artists tend to use in the titles.

What words appear more often in the lot titles?

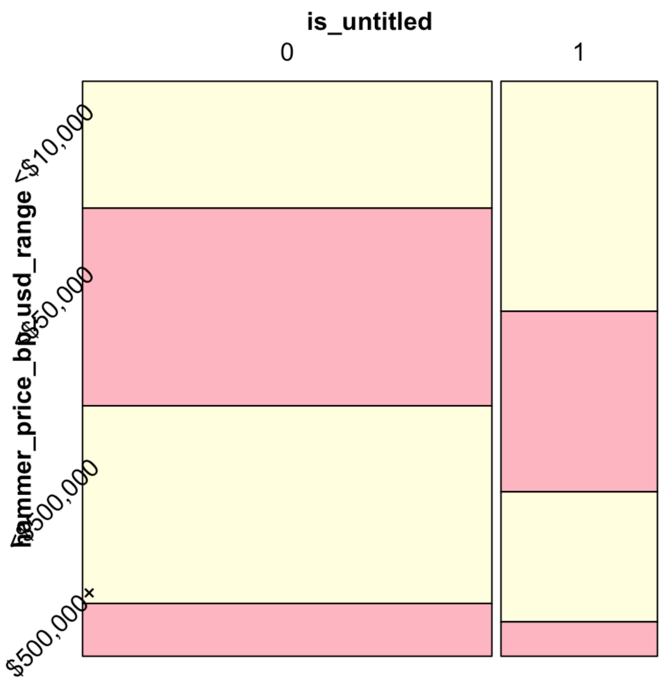
```
# collapse the lot_title column by word and count the frequency they appear in the titles.
temp <- paste(df1$lot_title, collapse=' ' )
temp <- tolower(temp)
temp <- gsub(" *\\b[[:alpha:]]{1}\\b *", " ", temp)
temp <- gsub('[[:punct:] ]+',' ',temp)
temp <- as.list(strsplit(temp, " "))
temp <- unlist(temp)[!(unlist(temp) %in% stopwords("english"))]
temp <- unlist(temp)[!(unlist(temp) %in% "na")]
word_count <- na.omit(as.data.frame(table(temp)))
word_count <- arrange(word_count,desc(word_count$Freq))[1:300,]

# visualize word frequencies
pal <- brewer.pal(9, "Dark2")
wordcloud(word_count$temp, word_count$Freq, min.freq =20, scale=c(5, .5), random.order = FALSE, random.color = FA
LSE, colors= pal)
```



Without surprise, “untitled” indeed are the most popular word artists tend to use. In that case, are “untitled” works more likely to receive higher prices?

Looking at lots that have name “Untitled”, what price ranges are they in? Is it correlated?

```
library(vcd)
df1 <- df1 %>%
  dplyr::mutate(hammer_price_bp_usd_range = forcats::fct_relevel(hammer_price_bp_usd_range, "<$50K"))%>%
  dplyr::mutate(hammer_price_bp_usd_range = forcats::fct_relevel(hammer_price_bp_usd_range, "<$500K"))%>%
   dplyr::mutate(hammer_price_bp_usd_range = forcats::fct_relevel(hammer_price_bp_usd_range, "$500K+"))

vcd::mosaic(hammer_price_bp_usd_range~is_untitled, direction = c("v", "h"),df1,
       gp = gpar(fill = c("lightyellow", "lightpink")),
                 labeling = labeling_border(rot_labels = c(0, 45),pos_labels="center"))
```

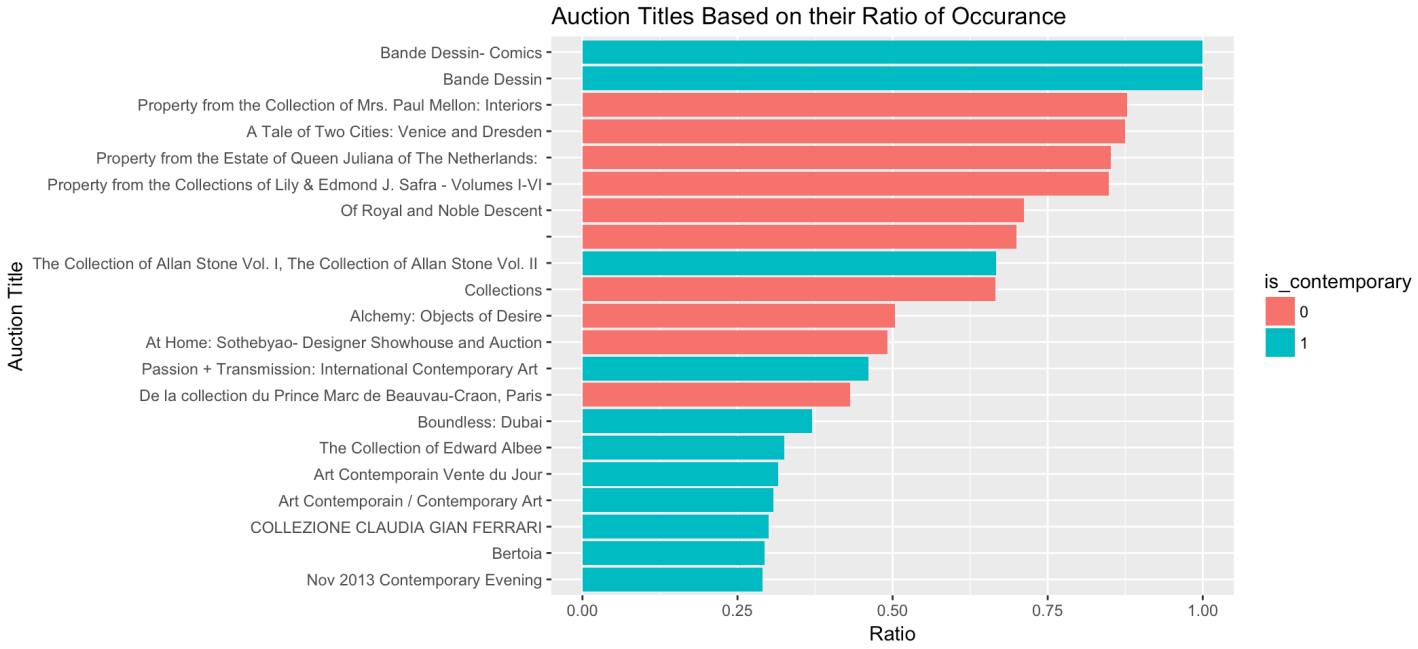


In the above plot, “1” indicates the lot has “untitled” in it and “0” otherwise. Analyzing the mosaic plot above, we can see that “untitled” actually result in a relatively lower price as fewer “untitled” lots are in the range “$500,000+” while more “untitled” lots are in the lowest range. If the higher prices cannot explain the common use of “untitled”, what can?

Next, we hypothesize that contemporary artists tend to name their arts “untitled”. To explore this, we put together the auction titles where “untitled” works are presented the most and labeled them with art type.

Are “untitled” works mostly contemporary?

```
untitle_ratio = read.csv("untitle_ratio.csv", header=TRUE)
#tempstr <-  strtrim(untitle_ratio$auc_title, 20)
#untitle_ratio$auc_title <- tempstr
untitle_ratio$is_contemporary = as.factor(untitle_ratio$is_contemporary)
ggplot(data=untitle_ratio, aes(x=reorder(auc_title,ratio), y=ratio,fill=is_contemporary))+geom_bar(stat="identity
")+coord_flip() + xlab("Auction Title") + ylab("Ratio") +ggtitle("Auction Titles Based on their Ratio of Occuranc
e")
```

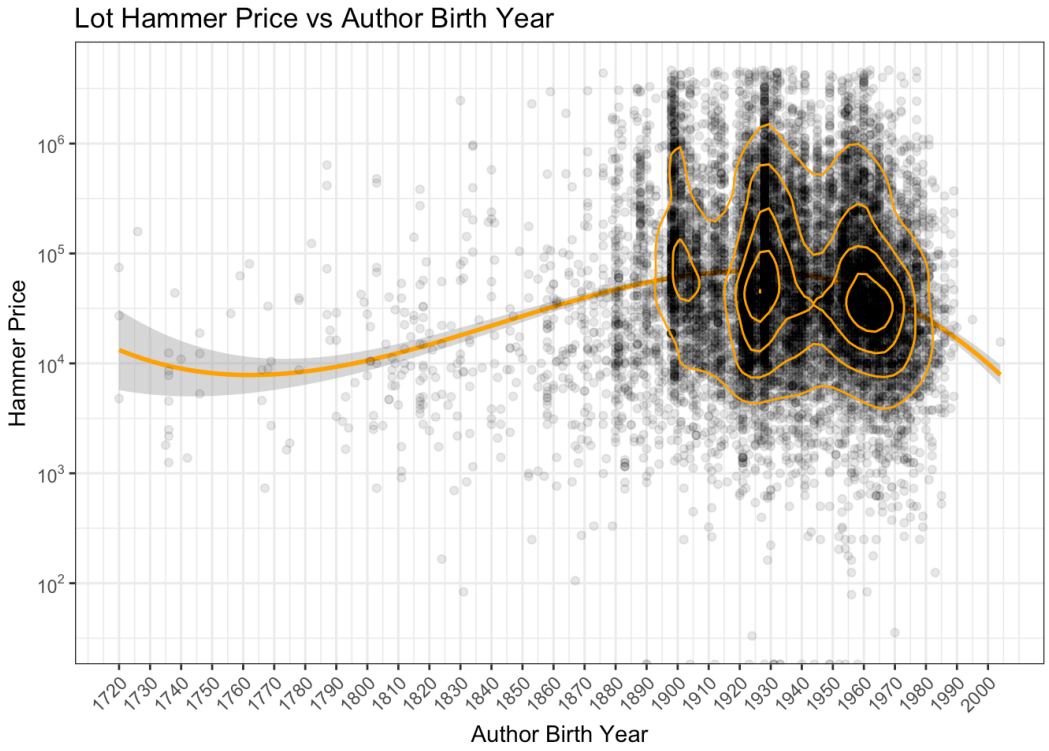


Of the top 20 auctions that have the highest “untitled” ratio, we see that contemporary auctions are just slightly greater than non-contemporary auctions. In fact, by looking at auction names of non-contemporary auctions, we realize that smaller pieces of art (such as decoration, furnitures) tend to have “untitled” in their title as well! Perhaps those contribute mostly to the fact that “untitled” work have relatively lower prices.

Does the era of the lot affect its price?

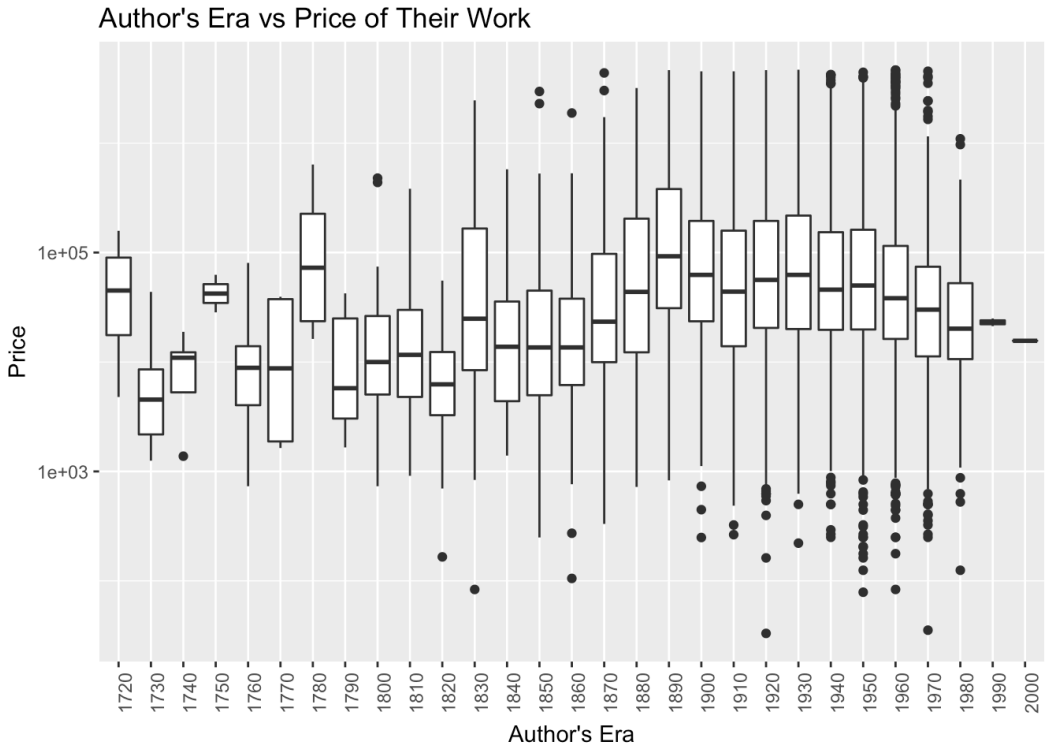
Here, we will plot the hammer price against artists’ birth years. Note that we are using log scales on the y-axis as the range of prices is wide.

```
df3 <- df1 %>%
  filter(df1$birth_year>1700)
ggplot(df3, aes(birth_year,hammer_price_bp_usd))  +
  geom_smooth(method='lm',formula=y ~ poly(x, 3),color="orange")+
  geom_point(alpha = .1)  +
  theme_bw()+scale_y_log10(breaks = scales::trans_breaks("log10", function(x) 10^x),
   labels = scales::trans_format("log10", scales::math_format(10^.x)))+geom_density_2d(bins = 5, color="orange")+
    scale_x_continuous(breaks = seq(min(df3$birth_year), max(df3$birth_year),10))+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  xlab("Author Birth Year")+
  ylab("Hammer Price")+
  ggtitle("Lot Hammer Price vs Author Birth Year")
```



From the above chart, we can see that modern pieces have larger variance. We'd like to try a box plot to capture this feature.

```
#box plot price vs year
ggplot(df3, aes(auth_era, hammer_price_bp_usd)) +
  geom_boxplot()+theme(axis.text.x = element_text(angle = 90, hjust = 1))+scale_y_log10() +xlab("Author's Era") +
ylab("Price") +ggtitle("Author's Era vs Price of Their Work")
```



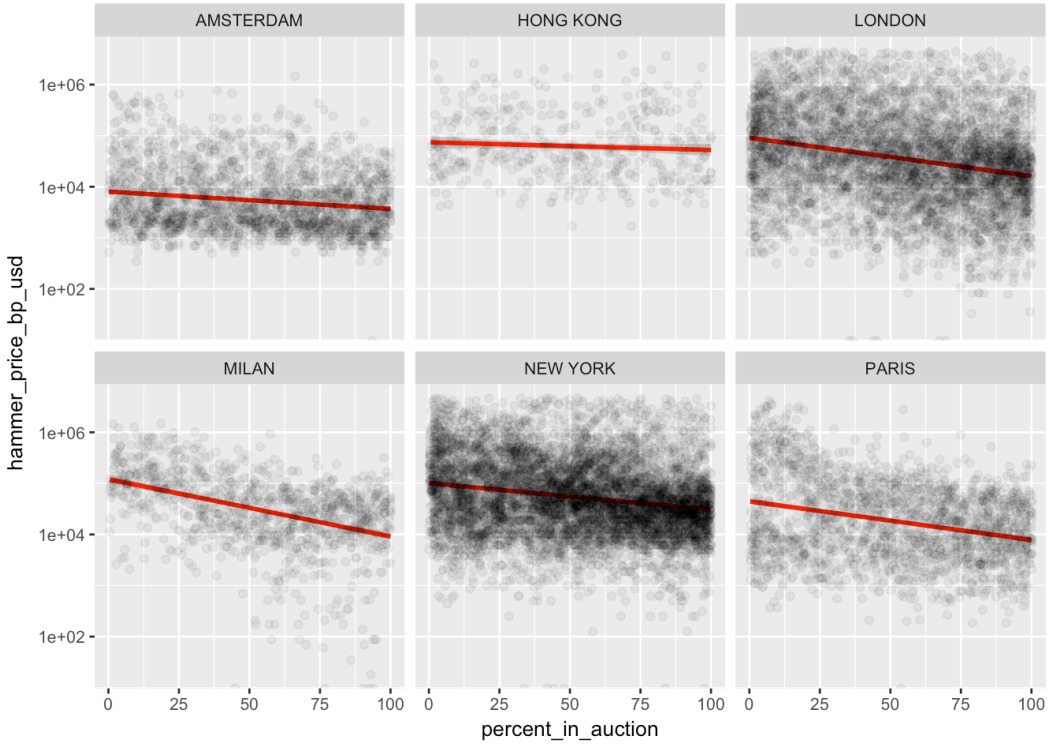
We notice that for certain eras (e.g. 1750), there are no pieces sold. There are also more outliers for authors born 1900 - 1980. However, due to the complex change in average price over every 10 years, this box plot looks messy overall. Therefore, we decided to only include the scatter plot as that contains the information more straightforward.

Do Certain External Factors Result in Higher Price?

Does the Order Matter?

In general, there are many lots in each auction. In our dataset, the average number of lots per auction is 357. With the large amount of lots being auctioned, we assume that the most valuable pieces get presented early in the auction. To validate this idea, we normalized the order of lots presented in each auction and plot it against the hammer price.

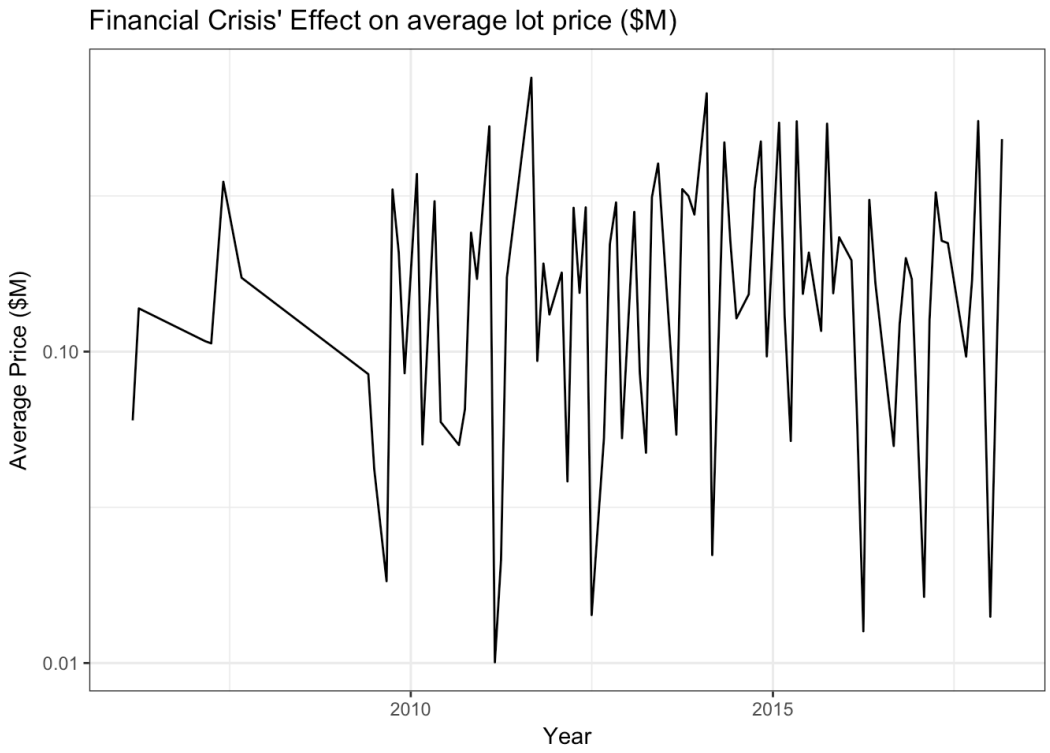
```
ggplot(df1, aes(percent_in_auction,hammer_price_bp_usd))  +
  geom_smooth(method='lm',formula=y~x,color="red")+
  geom_point(alpha = .05)  +
  theme_grey(10)+scale_y_log10()+
  facet_wrap(~location)
```



Is there an impact from the financial crisis?

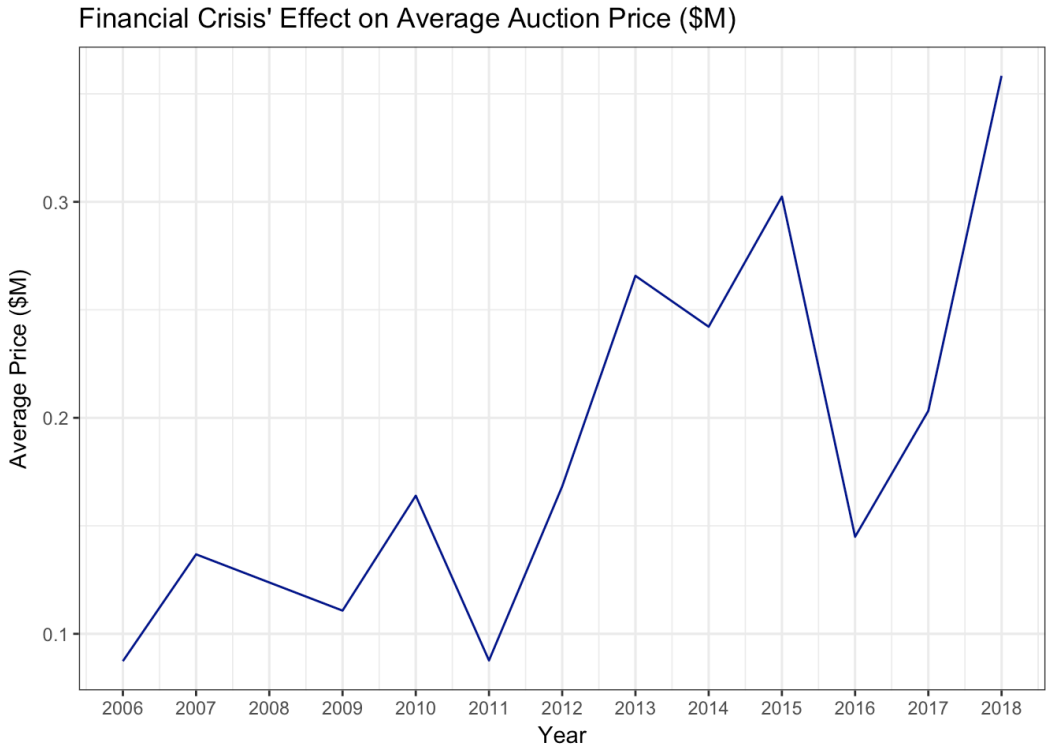
Let's start by looking at the average lot prices of Sotheby's on a monthly scale.

```
df1$auc_ymd <- as.Date(df1$auc_year_month_date)
art_yearfin <- df1 %>% group_by(month=lubridate::floor_date(auc_ymd, "month")) %>% summarise(revenue = mean(hamme
r_price_bp_usd))
ggplot() +geom_line(data=art_yearfin, aes(x=month, y=revenue/1000000))+ggtitle("Financial Crisis' Effect on avera
ge lot price ($M)")+ylab("Average Price ($M)")+xlab("Year")+theme(axis.text.x = element_text(angle = 45, hjust =
1))+scale_y_log10() +theme_bw()
```



It appears that the monthly scale is not very helpful in determining the effect because of the different fluctuations around each auction. Perhaps looking at the average price with a little less granularity will help? Since the monthly scale appears to have too much “noise,” we redrew the graph looking at a yearly scale instead. Our guess was that we are supposed to see a significant drop around the time of the financial crisis.

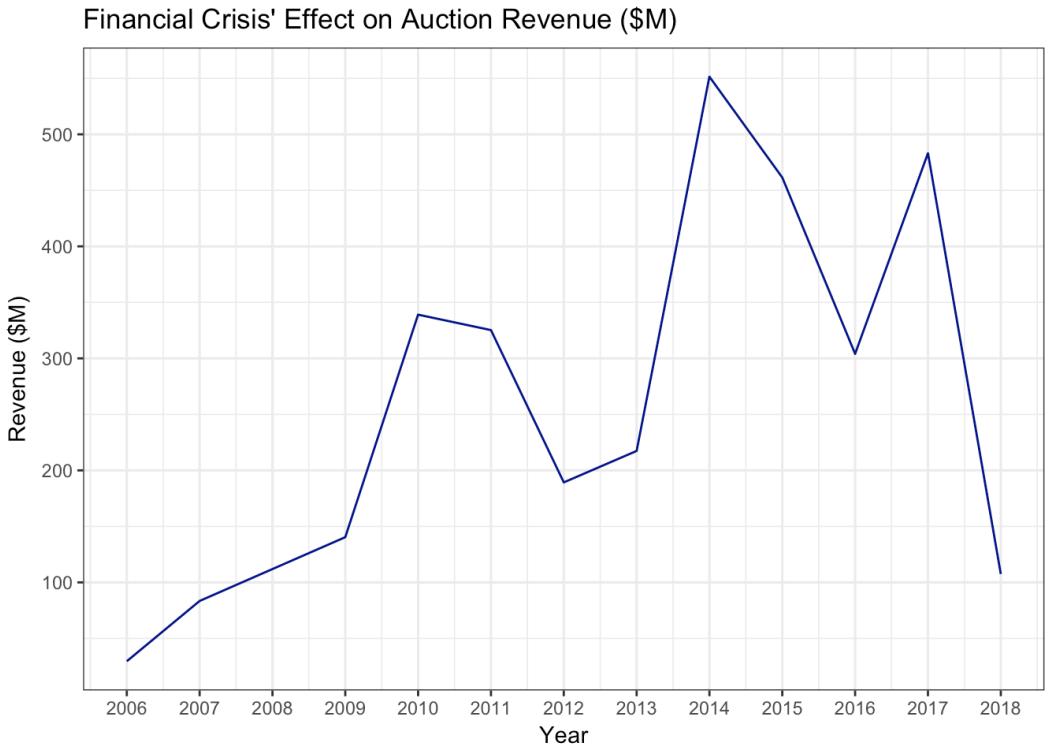
```
art_finance <- art_final[c("auc_year", "auc_month", "location","hammer_price_bp_usd")] %>% filter(!is.na(hammer_p
rice_bp_usd))
art_yearfin <- art_finance %>% group_by(auc_year) %>% summarise(av_revenue =mean(hammer_price_bp_usd))
ggplot(art_yearfin, aes(x=auc_year, y=as.numeric(av_revenue/1000000))) +geom_line(col="darkblue")+ggtitle("Financ
ial Crisis' Effect on Average Auction Price ($M)")+ylab("Average Price ($M)")+xlab("Year") +theme_bw()+scale_x_co
ntinuous(breaks = seq(min(art_final$auc_year), max(art_final$auc_year),1))
```



Indeed, we observe a big dip around late 2010. It may be surprising to see that it took some time for the effect to reach the auction houses. Nonetheless, when thinking back on the global time line of the Financial Crisis' effect, it did take a few years for it to reach other industries outside of finance.

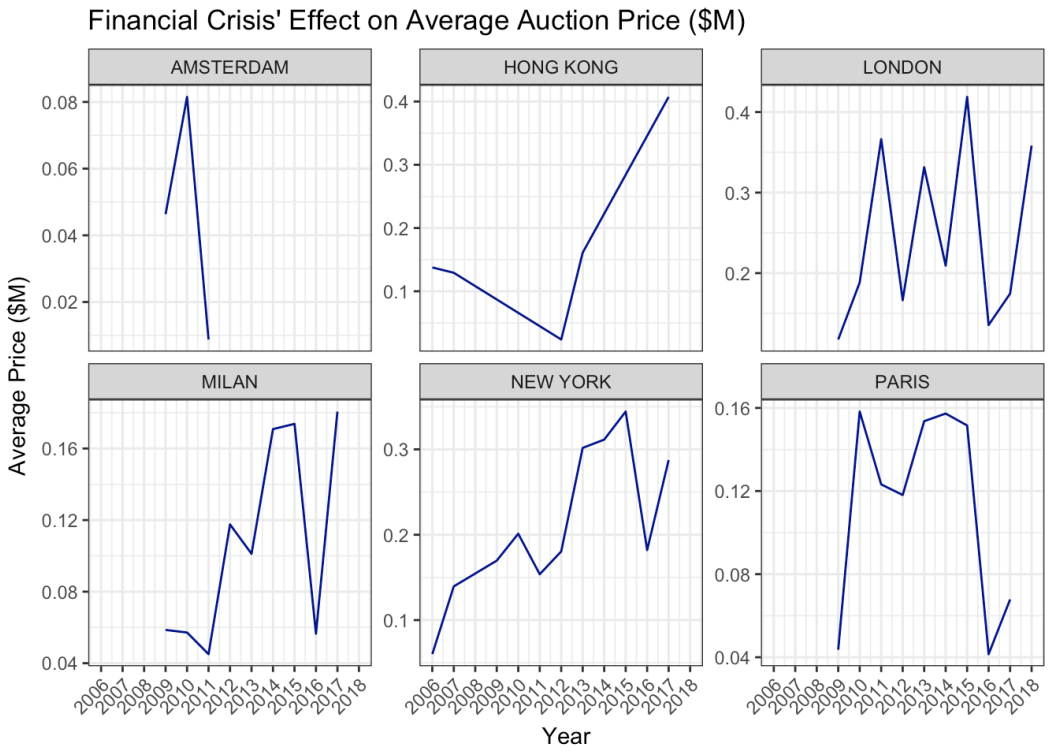
Next, let us explore the same effect on the total revenue and see if the same pattern holds.

```
art_finance <- art_final[c("auc_year", "auc_month", "location","hammer_price_bp_usd")] %>% filter(!is.na(hammer_p
rice_bp_usd))
art_yearfin <- art_finance %>% group_by(auc_year) %>% summarise(revenue = sum(hammer_price_bp_usd))
ggplot(art_yearfin, aes(x=auc_year, y=as.numeric(revenue/1000000))) +geom_line(col="darkblue")+ggtitle("Financial
Crisis' Effect on Auction Revenue ($M)")+ylab("Revenue ($M)")+xlab("Year")+theme_bw()+scale_x_continuous(breaks =
seq(min(art_final$auc_year), max(art_final$auc_year),1))
```



We can easily tell that the two graphs are very similar and that total revenue did experience the same drop. Perhaps there are certain locations that are skewing some of the data. Let us try to facet the data by location and see if that could present us with a better outlook. We will be excluding Dubai and Doha due to lack of data for hammer prices (as shows in the data quality part).

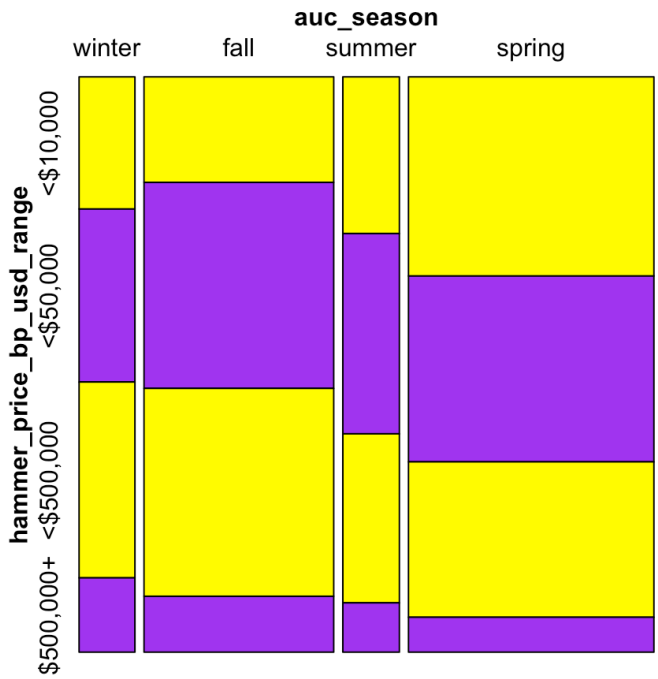
```
art_locfin <- art_finance %>% group_by(auc_year, location) %>% summarise(av_revenue = mean(hammer_price_bp_usd))
%>% filter(location %in%  c("AMSTERDAM","HONG KONG","NEW YORK", "LONDON", "PARIS", "MILAN"))
ggplot(art_locfin, aes(x=auc_year, y=as.numeric(av_revenue/1000000))) +geom_line(col="darkblue")+ggtitle("Financi
al Crisis' Effect on Average Auction Price ($M)")+ylab("Average Price ($M)")+xlab("Year") + facet_wrap(~location,
scales ="free_y")+theme_bw()+scale_x_continuous(breaks = seq(min(art_final$auc_year), max(art_final$auc_year),1))
+theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We notice a consistent drop in revenue across all locations above starting 2010. Therefore, our original hypothesis seems correct: the financial crisis did have an effect on the auction revenue across the world (specifically significant drops are observed in New York and Hong Kong).

Does Season Matter?

```
df1 <- df1 %>%
  dplyr::mutate(auc_season = forcats::fct_relevel(auc_season, "summer")) %>%
  dplyr::mutate(auc_season = forcats::fct_relevel(auc_season, "fall")) %>%
  dplyr::mutate(auc_season = forcats::fct_relevel(auc_season, "winter"))
vcd::mosaic(hammer_price_bp_usd_range~auc_season, direction = c("v", "h"),df1,
       gp = gpar(fill = c("yellow", "purple")),
                 labeling = labeling_border(rot_labels = c(0, 90),pos_labels="center"))
```

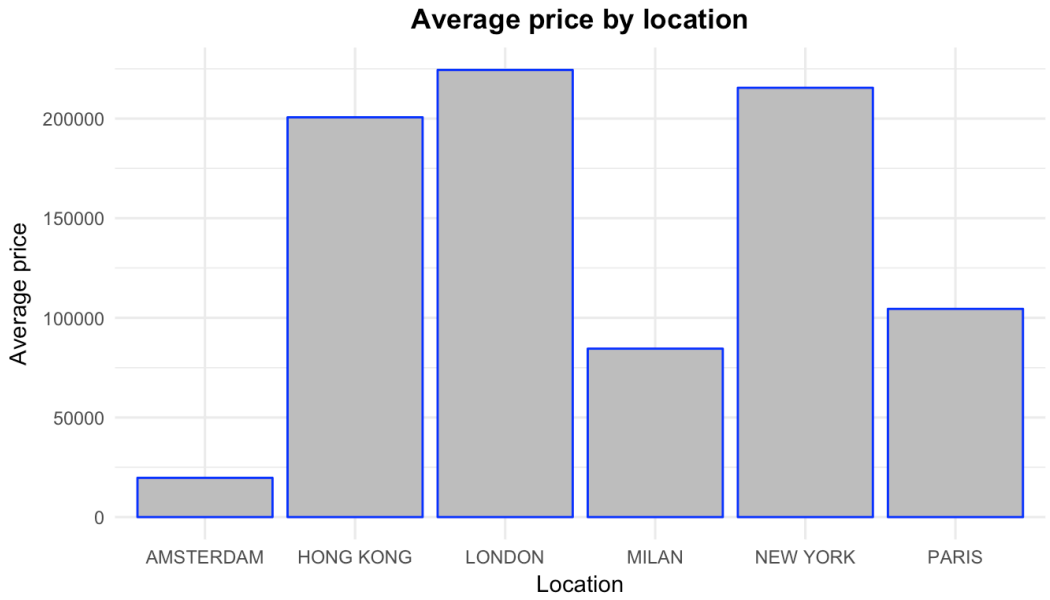


The labels of hammer price range did not turn out as neat in the beginning, however, after changing the order of the seasons, we obtained the above graph indicating the clear correlations between seasons and the lot prices. Specifically, there are more lots sold in spring and fall and lots are sold relatively higher in the fall.

Does Location Matter?

To explore which locations have higher average price we created a bar chart.

```
MyData_2a <- subset(art_final, select=c( "location", "hammer_price_bp_usd" ))


MyData_2b <- MyData_2a %>% group_by(location)%>% summarise(B=mean(hammer_price_bp_usd))

ggplot(MyData_2b, aes(x= location, y = B)) +
    geom_bar(stat='identity', color="blue", fill="grey")+labs(y = "Average price")+labs(x = "Location") + ggtitle
("Average price by location") + theme_minimal()  + theme(plot.title = element_text(hjust = 0.5,face="bold"))
```



From the graph it can be seen that Hong Kong, London and New York have the highest average lot price. Which confirms our assumption from part 1 that New York and London are the main hubs for art exchange.

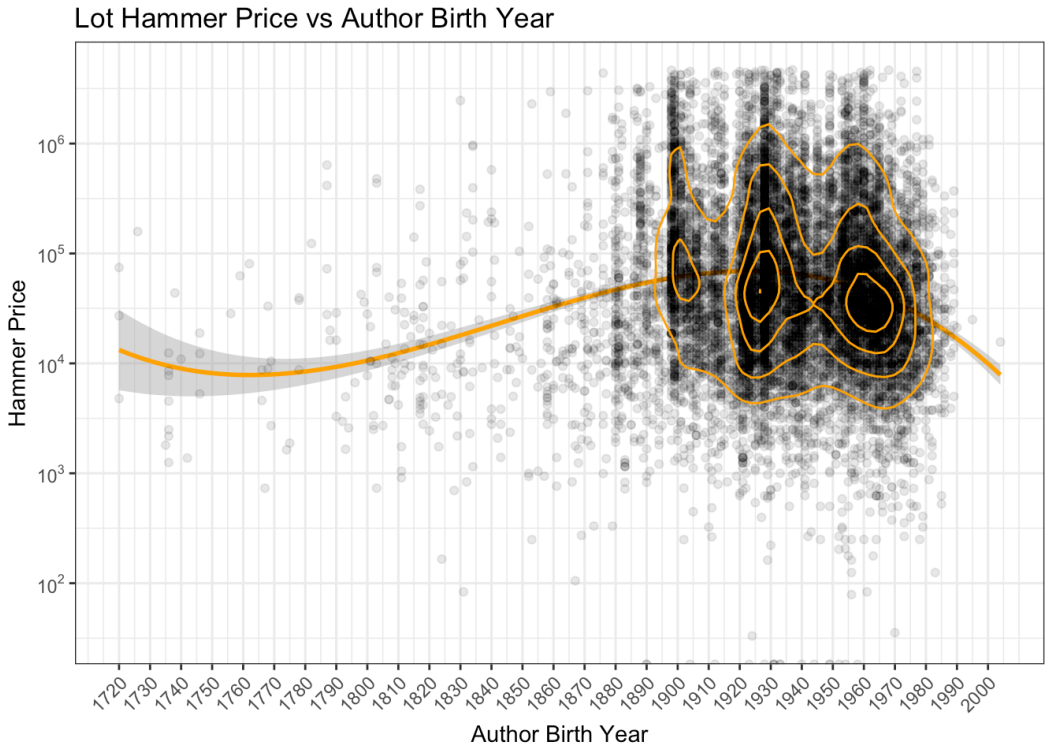
Executive Summary

The explanatory visualization part of the project allowed us to dive deep around our original questions of interest. We used the main analysis to narrow down some of the most interesting findings in relation to our dataset.

Next, we will present these findings with graphs in following summary:

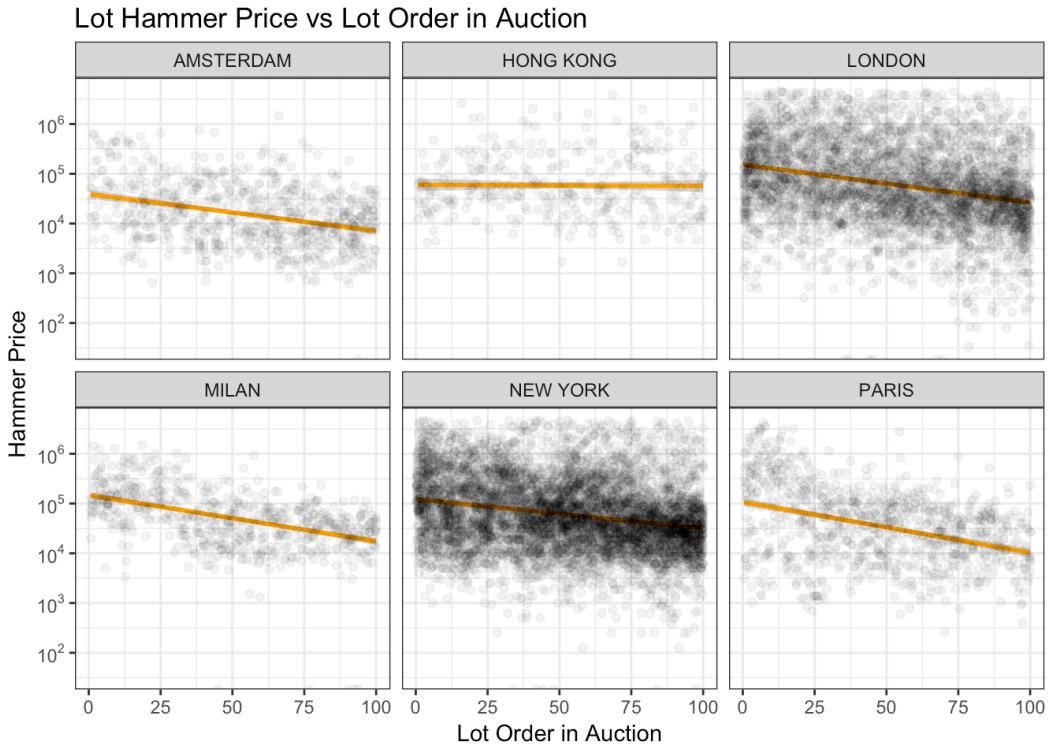
1. Art pieces created in the 20th century on average are sold at a higher price.

There are more lots sold where their authors are born in the 20th century, specifically in 1900s,1930s and 1960s. For simplicity, we call them “modern” pieces and call the lots whose authors were born before 1900 “older” pieces. From the graph, we can see that modern pieces have higher prices on average while having larger variance too. One plausible explanation is that the price range and values for contemporary art are more likely to uneven due to the nature of the contemporary art.



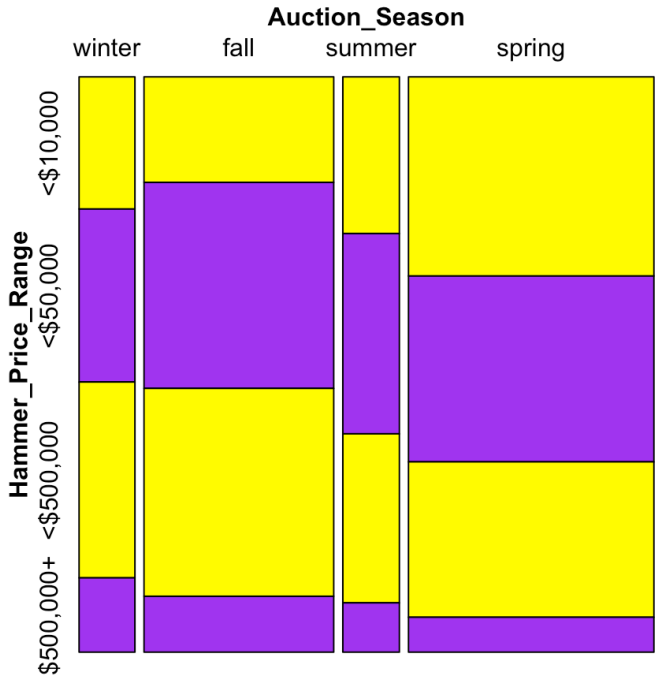
2. The earlier the lots get presented in an auction, the higher price they have.

Due to large number of lots presented in Sotheby's auctions, the most valuable pieces get presented early on in the auction. From the below graph, we conclude that there is a negative relationship between art prices and order presented for all locations besides Dubai. Particularly Paris has the most obvious negative relationship.



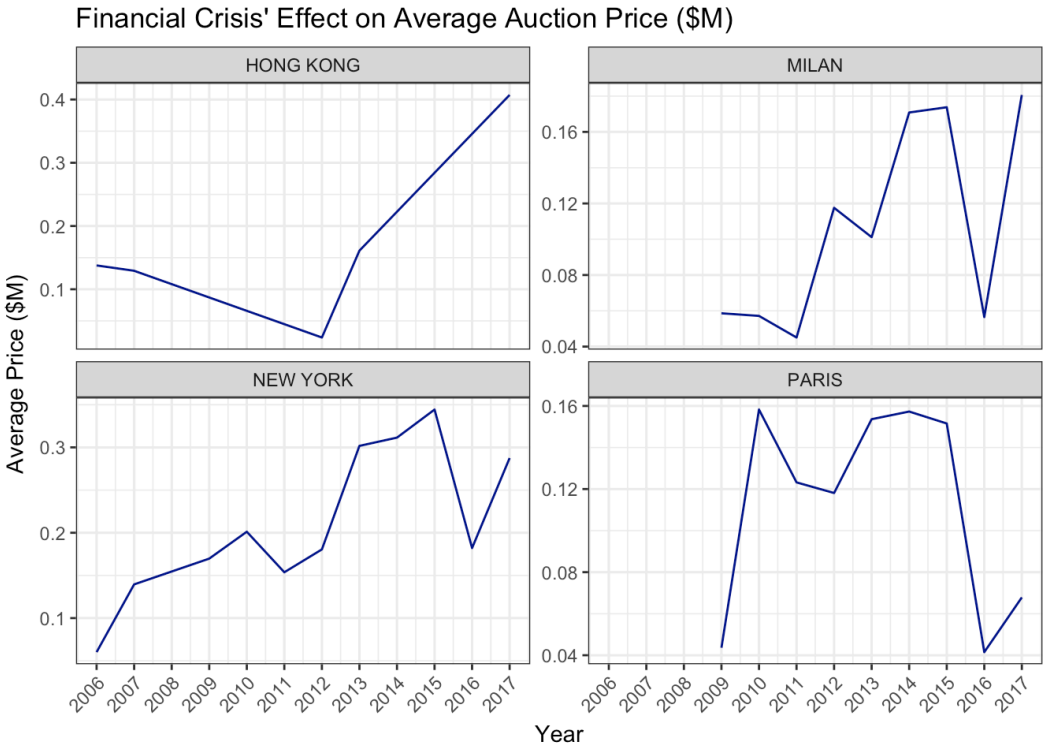
3. Lot prices are closely related to season.

Most of the auctions appear in March and November, thus in the graph we see fall and spring have more lots being sold. Aside from the quantity, the price of the lots are closely related to season as well. We conclude that winter has a positive correlation for lots to be sold in the highest range ($500,000+) as well as the lowest range (<$10,000). Lots sold in Fall, however, appear more likely to fall in the higher price ranges. Lots sold in spring are relatively cheaper compared to lots sold in other seasons.



4. Financial Crisis had a negative effect on average price across almost all locations around the world

By looking at the graph, one can see how each location experienced a significant drop in average price approximately 2-3 years after the initial Financial Crisis hit in 2008. While some locations have a stronger and clearer trend than other, the directional conclusion stays the same. Some locations like New York and London lost close or even more than 50% of value (average price) during that time.



In addition, it is important to notice the major climb in prices around 2012-2013. This observation goes in support with our original hypothesis that investors began putting more money into art as a new form of investment after the financial crisis. In all locations (with an exception of Paris), the average lot price has recovered and surpassed the one before the crisis.

Interactive Component

Link to the interactive part: https://edav-art-viz.firebaseapp.com (https://edav-art-viz.firebaseapp.com)

Description

Our interactive visualization is a scatter plot of lot prices over time built with HTML, CSS, Bootstrap and JavaScript; it is hosted on Firebase and can be viewed here: https://edav-art-viz.firebaseapp.com/ (https://edav-art-viz.firebaseapp.com/)

The visualization allows users to see all of the lots sold according to their price and locations each year. Each color represents a different location, and the black horizontal average line indicates the average hammer price for all lots during the selected year. This visualization supports the project's main hypothesis that location and season significantly affect the auction performance in terms of the prices of works sold.

We built the interactive visualization so that it would be scalable and versatile if we want to add more features in the future. A few things that we had in mind as features but did not implement due to time constraints included taking the average per auction, and connecting these points with a trend line. We also thought of adding a box plot for each auction that could have been toggled in a similar matter to the jitter feature. Another feature we had hoped to add was the ability to toggle based on location, which would have allowed users to click each location on or off and show the corresponding lot data points. Finally, our time line only allows for scrolling by year, but ultimately the idea would be to have a continuous scroll and even zoom in and out so the user would have full control over their time frame of analysis.

Lastly, both for the interactive visualization and our analysis on a whole, future work would include extending our project to not only analyze auction data from Sotheby's, but also to analyze data from other auction houses like Christie's and Phillip,s or museums and other online art data sources to see if our conclusions are consistent for these other data sources as well.

Instructions

Each vertical line indicates an auction, and all the data points along that line are the lots (pieces of art) that were sold in that auction. Auctions are colored by location so users can see where in the world the auction took place. The average line indicates the average price at which lots were sold in aggregate for the selected year. Additionally, there are a number of ways users can interact with this scatter plot including:

1. **Change Year** - Click on the arrows next to the year to either move a year forward or backwards (2006-2018). This can help to compare the different distributions and average price of lots across all years in our dataset.
2. **See Lot Details** - Hover over each circle to see the details of the lot transaction. You should expect to see the realized hammer price (sale price) as well as a picture of the actual piece of art. Keep in mind that many of the lots have their pictures copyright protected by Sothebys and will not be display. Nonetheless, most of the more recent images (2017-18) are available. :)
3. **Spread Data Points** - Since there are so many data points per auction plotted along the same day, users can add jitter to better observe each data point and counteract overplotting. Click the button that says "Jitter" and this will spread the points from their original location to better see all options. One can easily bring the points back to proper location by unclicking the button.

Conclusion

Through our exploration with Sotheby's auction data, we not only learned interesting trends in the art market and understood how different factors affect lot prices (such as order presented in auction, impact of financial crisis etc), but also gained a deeper understanding in visualization theory and practiced interactive visualization with modern technology. Some of the lessons learned are:

1.Scaling is important. At first, we used continuous scale in our y-axis when it comes to plotting prices. However, due to large range and uneven distribution of the price, the graphs gave us - trivial information. Therefore, we decided to use log scale for most of our analysis. It is worth to note that adopting log scale has the limitation to label meaningfully as it would be hard to show "millions" in a log scale.

2.Data format and quality is essential to any good analysis. Without excluding NAs and filtering some of the locations with missing data we would not have been able to successfully conduct the analysis.

3.Defining questions before starting the analysis makes the work more efficient and productive. The questions guided us through the process and allows us to stay on track.

4.Preprocessing is key to deep analysis. Most of our visualizations are actually built on the features (columns) that were artificially created through data preprocessing. Anticipating analysis needs and creating those extra columns allows us to go above and beyond in answering the questions asked.

When conducting the analysis, there were limitations due to our data. For example, when analysing the effect of certain lot titles (e.g. "untitled") on lot prices, we weren't able to distinguish contemporary auctions from other types of auctions on a larger scale. Thus, if provided auction type field, we can explore this area further.

For future analysis, we'd also be interested in collecting more data from other auction houses such as Christie's and analyzing trends on a greater scale. In addition, we are curious if the findings are different between different auction houses and if so, why. It might be also interesting to explore the trends in the art market for Asia, since in our graphs, Hong Kong has a high average price among different locations. The rising interest towards art auctions was widely covered in the different art publications. Lastly we could further explore is the amount of unsold lots for every auction. This data was not included in our data set, but can be added based on Sotheby's resources and publications.