Final Project

Shreya Chaganti Mert Ketenci Victor Allan

10 December 2018

```
library(tidyverse)
library(gridExtra)
library(reshape2)
library(boot)
library(purrr)
library(dplyr)
library(data.table)
library(extracat)
library(lubridate)
library(RColorBrewer)
library(ggmosaic)
library(choroplethrZip)
library(tidyquant)
library(dplyr)
library(choroplethr)
library(scales)
library(ggthemes)
art=read.csv('MOMA_art_cleaned2.csv')
```

1. Introduction

Our initial interest in pursuing this area of exploration was to learn more about art and whether we could find any features that would help explain what made a piece of art valuable. After doing some preliminary research, we decided to focus on data from the Museum of Modern Art in New York. We reasoned that if a piece was in a prestigious museum such as MoMA, it would be considered valuable. Additionally, we found that a large proportion of the most expensive pieces of art in history were modern and contemporary art. We decided to analyze the artworks in the MoMA to explore their characteristics and see if there were any general trends within the collection. The data we obtained contains information regarding to art pieces in the New York Museum of Modern Art. There are lots of interesting questions one can ask about art, and a lot of insight we can gain by analyzing the pieces in the MoMA. Some of the questions we will be addressing are as follows:

Artists in General:
-The Most Productive Age of a Certain Artist -What is the most productive age for artists in general?
Gender:
-Are there any patterns for productivity with respect to gender by time?
Nationality:
-Which nationality expresses themselves with what type of art? (normalize with respect to nationality)
Patterns in museum:
-Are there any patterns between missing data and categorization?
-Are there any patterns for acquiring art? (Yearly-monthly-seasonality)
-Does the demand change after the artist dies?
-Are there any patterns in terms of how the piece was acquired (gift, purchased, etc)?
-Are there any shared characteristics (in terms of productivity) of the top 10 highest selling artists?
Geographic Patterns:
-From which countries artworks came from?

In this project, each team member wanted to be involved in every aspect of the project rather than dividing the work. Thus, we all worked on each EDA, cleaning data and preparing project equally and met on a regular basis to do so.

2. Description of Data

The source of the data is data.world, which is a platform where institutions can share data. The data we are going to analyze is shared by New York, Museum of Modern Art. The page of New York, Museum of Modern Art on data.world is https://data.world/moma. The data is also available on the museum's Github page (https://github.com/MuseumofModernArt/collection). The attributes we are going to use in this project and their description are as follows:

1.tittle : The tittle of the art piece
2.artist : The name of the artist
3.constituentid : The id given to art piece
4.artistbio : Combination of nationality and birthday
5.nationality : The nationality of the artist
6.begindate : Birthday of artist
7.enddate : The day artist passed away
8.gender : The gender of the artist
9.date : The date art piece was completed
10.medium : The medium of the composition
11.dimensions : The dimensions of the piece
12.creditline : The donation/gift credit
13.classification : The type of the art piece
14.department : The department where the piece is exhibited
15.dateacquired : The date when the art piece acquired
16.cataloged : Indicator if the piece is catalogued
17.circumference_cm : Circumference of the piece
18.depth_cm : Depth of the piece
19.diameter_cm : Diameter of the piece
20.height_cm : Height of the piece
21.length_cm : Length of the piece
22.weight_kg : Weight of the piece
23.width_cm : Width of the piece

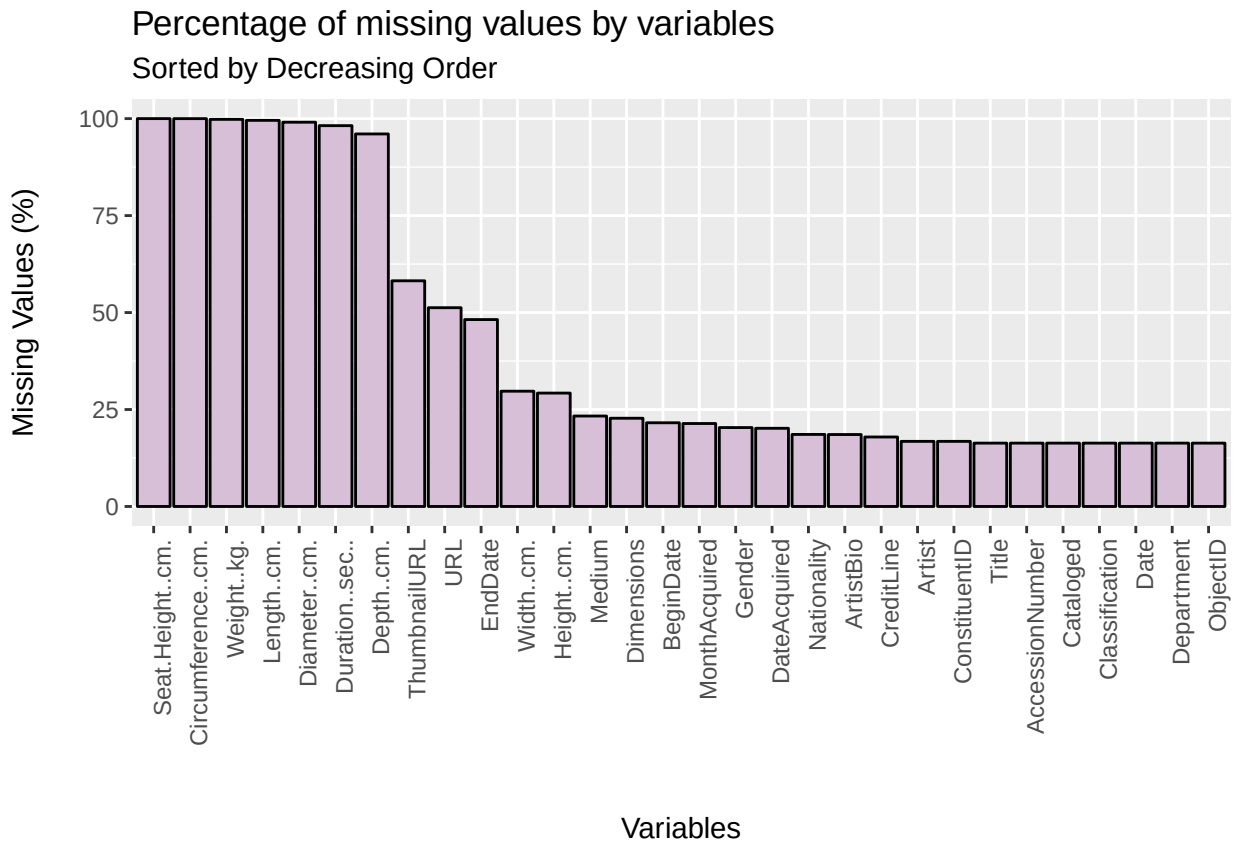
3. The Analysis of Data Quality

In this section we are going to draw a bar plot to observe the quality of the data. The plot is going to be based on the percentage of missing values for each column. As the variables that are missing most are geometric attributes of the pieces, we can conclude that the Museum of Modern Art does not keep track of the geometric attibutes of many pieces or does not share that information for some pieces. There may also be pieces such as sculptures that may not have traditional measurements. It is important to consider that not all of the art pieces may have such attributes such as design, musical and architectural ones.

Fortunately, the variables that interest us the most are not missing.

```
art$X <- NULL
art$X0 <- NULL
art$Unnamed..0 <- NULL
#Let's see what are missing
missing_variables=colSums(is.na(art)) %>%
  sort(decreasing = TRUE)
#Let's turn what we have into a dataframe
```

```
missing_variables=data.frame(missing_variables)
#Let's put the row names as a seperate row to plot
invisible(setDT(missing_variables, keep.rownames = TRUE)[])
#Let's find the total before calculating the percentage
total=dim(art)[1]
#Convert into %
missing_variables$missing_variables=100*missing_variables$missing_variables/total
#Let's Create Labels
#Let's plot
#I am not going to put Label x name as it is obvious what it is
#Putting that would only make graph dense
ggplot(missing_variables, aes(x = reorder(rn,-missing_variables),
                              y = missing_variables)) +
  geom_col(color = "black", fill = "thistle") +
  ggtitle("Percentage of missing values by variables",
          subtitle="Sorted by Decreasing Order") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_discrete()+
  labs(x="\n\nVariables",y = "Missing Values (%)\n")
```

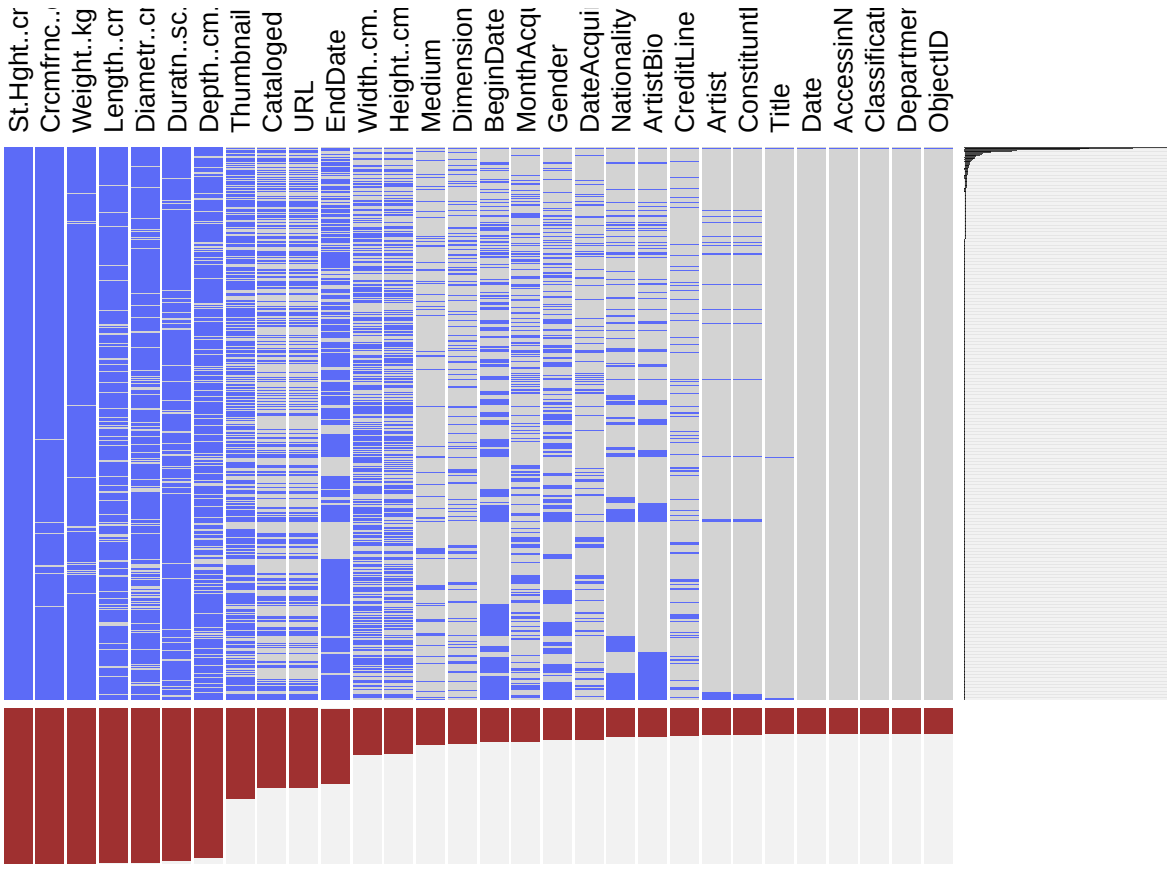


This is a good point to begin to seek the answer to the question **Are there any patterns between missing data and categorization?**. In other words, is there a pattern such that uncataloged pieces have more attributes missing than cataloged? To do this, we are going to replace the parts of cataloged column where cataloged = N with NA values and do a visna plot. There are also patterns where the piece is cataloged but certain attributes exist, though they are a minority.

In the below visna plot, we see that above all of the missing patterns of the variable cataloged, the most missing one is with also all geometrical variables missing. Thus, we can conclude that the measurements are taken when a piece is being cataloged.

We again can observe that the information regarding artists is not missing.

```
art$Cataloged[art$Cataloged == "N"] =NA
visna(art,sort = "b", mar.col = c(alpha("black", 0.7), alpha("darkred", 0.8)))
```



Challenges to Data Cleaning:

One of the biggest challenges in data cleaning was that the dates in the data were not in a clean format. Specifically the years were covered in parentheses and many of the dates were ranges or had text surrounding the year (i.e. “circa 1903”" or “not before 1942”). In order to solve these problems, we took only the beginning of the date range and we removed all text. We also removed all parentheses. One method we initially tried that did not work was to immediately run an r function to convert the column into a date object, before stripping the alphabetic characters. This method unfortunately did not work, and we had to fix the dates manually. Later in the analysis, we additionally converted the dates to DateTime objects.

Another issue we had with data cleaning was looking at nationality. Because some artists were born one place and then moved and became citizens of another, or an artist could have dual nationality, we tried to clean the data so that only one nation would show up or at least would not have a full description about the nationality (some nationalities were sentence long descriptions). We succeeded in removing the long descriptions and generally were able to take only the first nationality mentioned in the column.

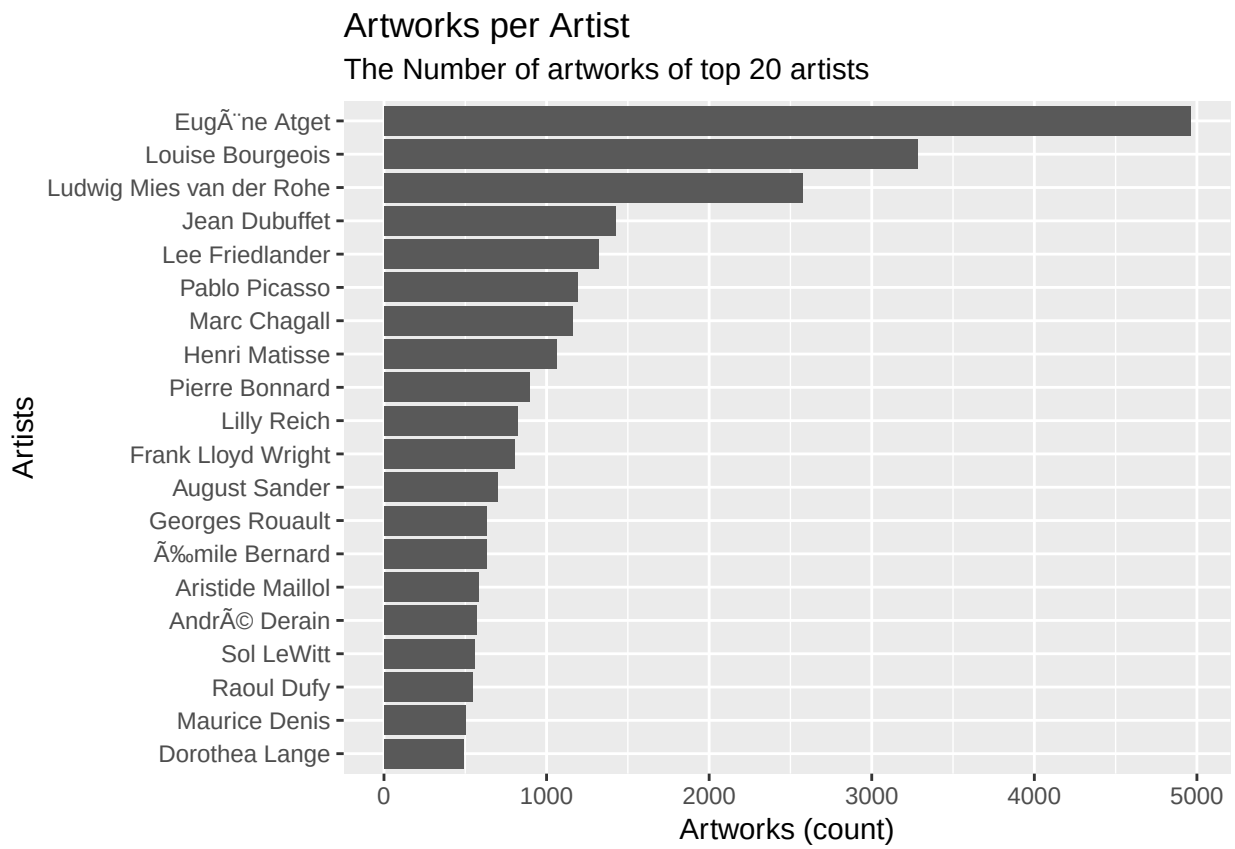
4. Exploratory Data Analysis

As we began our exploratory analysis, we wanted to get a sense of the overall data. We started by sorting various variables by frequency and creating bar plots to see which artists, mediums, classifications, etc. were the most common throughout the museum. One interesting aspect we found was that the bulk of the collection is not paintings, as one would think, but prints and illustrated books. After looking at the top artists by number of pieces, we noticed that Eugene Atget had an overwhelming amount of work - close to 5,000 pieces. Further exploration showed that Atget's pieces were almost entirely silver prints. Sorting by classification, or the category of art that a certain piece fell under, showed us that photographs, prints, and illustrated books were the categories with the highest quantities of pieces. Below is an example of one such exploratory histogram. Interestingly, there are only a few very well known names in the top 20 artists by volume, such as Picasso and Matisse. This can likely be explained by the larger effort it takes to create a painting versus a print, so it seems that the medium may also affect how prolific the artist was.

Something interesting we would have liked to do if given more time was color the fill of the bars by the proportion of classification or department of each artist's body of work, in order to get a visual sense of how each artist was broken down in terms of the type of art they created.

Following our initial exploration, we devised some questions that we thought we might be able to answer with the data.

```
artists <- art %>% filter(!is.na(Artist)) %>% filter(!(Artist == "Unknown photographer")) %>% filter(!(A
r2 <- artists %>% group_by(Artist) %>% summarise(Freq=n()) %>% arrange(desc(Freq))
top20 <- r2 %>% head(20)
ggplot(top20, aes(reorder(Artist, Freq), Freq)) + geom_col() + coord_flip() + ylab("Artworks (count)")+
  ggtitle("Artworks per Artist",
  subtitle = "The Number of artworks of top 20 artists")
```



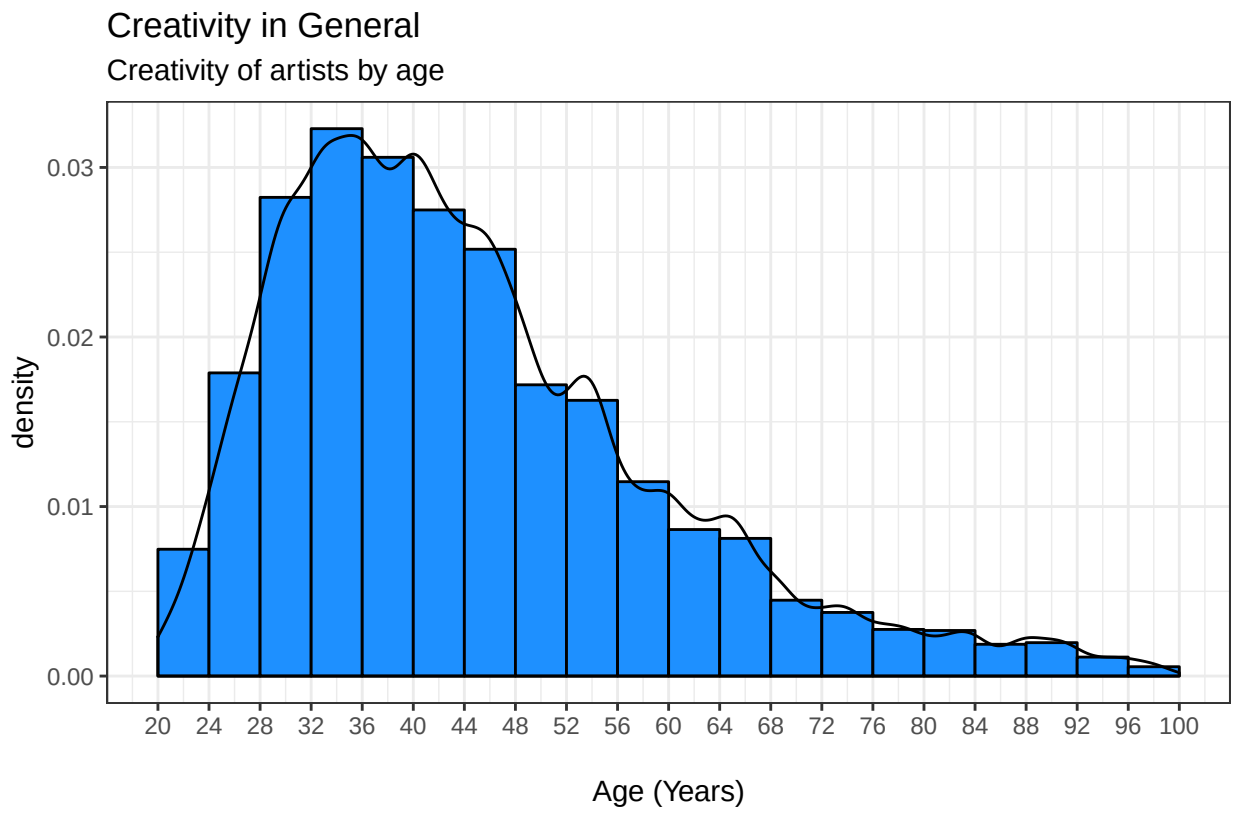
The most productive age of artists

In this section we are going to assume that works that are exhibited in New York Museum of Modern Art are the creative works of an artist. Thus, their count will represent how much creative work an artist produced. We will assume that the distribution of their work within the museum is an approximation of the distribution of their actual work. Therefore, we can consider their count to be an indicator of their creativity. As such, the count with respect to age is going to indicate the creativity of an artist with respect to his/her age. In this section we are going to create a new column for each artwork that indicates how old the artist was when he/she did the art work. At the below graph it is easy to observe that the most creative ages of an artist is between 30 and 42. The plot is a right skewed unimodal normal distribution.

It is interesting and important to state that our findings are in parallel with what the literature says. In his paper "When Did Classic Composers Make Their Best Work?" **Philip Hans Franses** from Econometric Institute, Erasmus School of Economics, Rotterdam, The Netherlands indicates that the peak creativity has been achieved when person is 38.9 years old. That is 41.9 for painters and 44.8 for authors.

```
art$Age = art$Date - art$BeginDate

ggplot(art, aes(x = Age, color=Age, y=..density..)) +
  theme_bw()+
  geom_histogram(bins=21, colour = "black", fill = "dodgerblue", boundary = 0)+
  ggtitle("Creativity in General",
          subtitle = "Creativity of artists by age") +
  labs(x = "\nAge (Years)", caption = "Source: world.data")+
  theme(plot.caption = element_text(color = "azure4"))+
  geom_density(alpha=0.2)+
  scale_x_continuous(breaks = seq(20, 100, 4), lim = c(20, 100))
```



Now let's look at the creativity of different artists in the same plot. For this plot, we are going to choose 3 artists: Pablo Picasso, Bernard Tschumi, and Vasily Kandinsky. On the below plot we observe 3 peaks for Vasily Kandinsky, which we can analyze further.

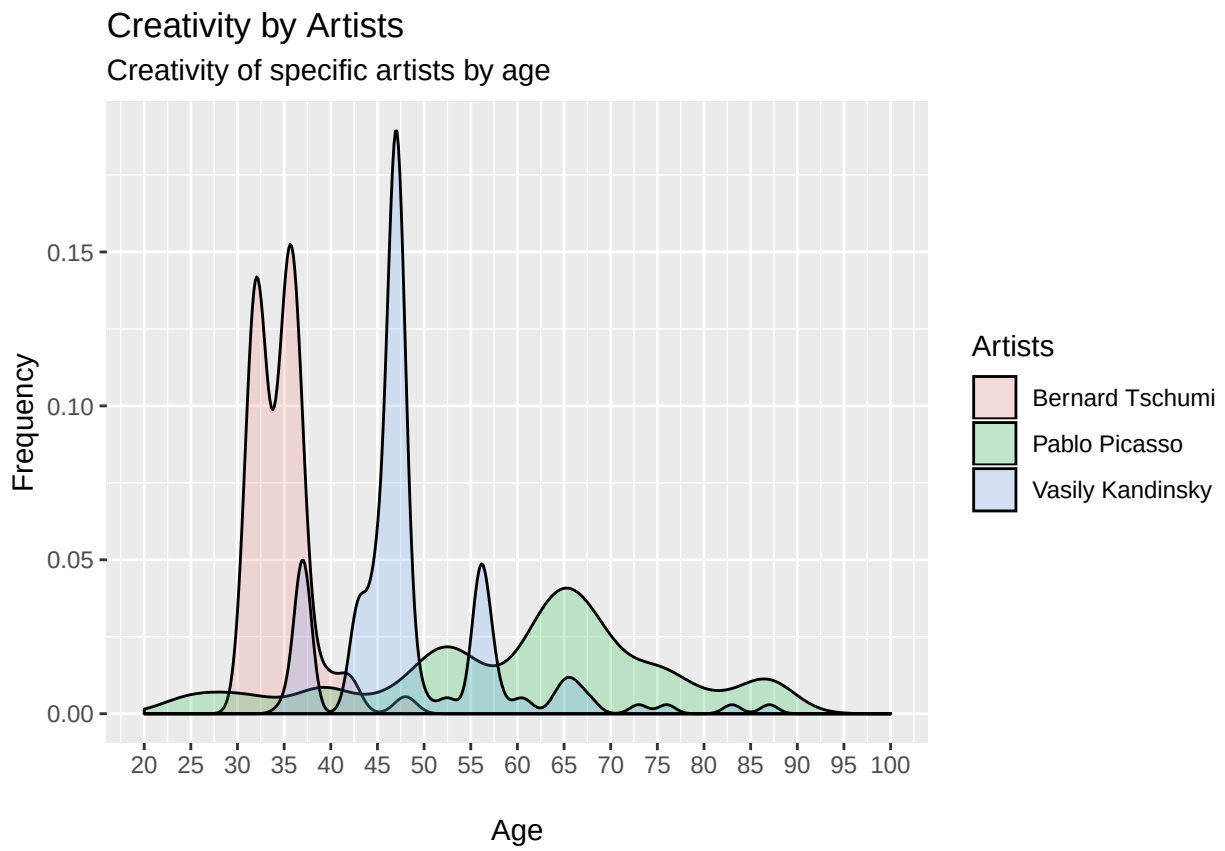
The first peak happens when he is 36 years old. That is the year 1902. That era corresponds to the time when Kandinsky began working with Gabriele Minter, called the metamorphosis era. Their relationship became personal rather than professional in the following years. Minter and Kandinsky's relationship affected Kandinsky's work and perhaps his creativity. Until the age of 40 we see a decline in his works.

At the age of 45, that is the year 1911, Kandinsky is in his Blue Rider Period (1911-1914). He reaches his peak. Famous paintings such as "Tableau a la tache rouge" belong to this era.

In July 1933. Kandinsky then left Germany, settling in Paris due to Nazi movements and lived there. In his older age, we can see the effects of the Second World War, as it is hard to do art works in such a situation.

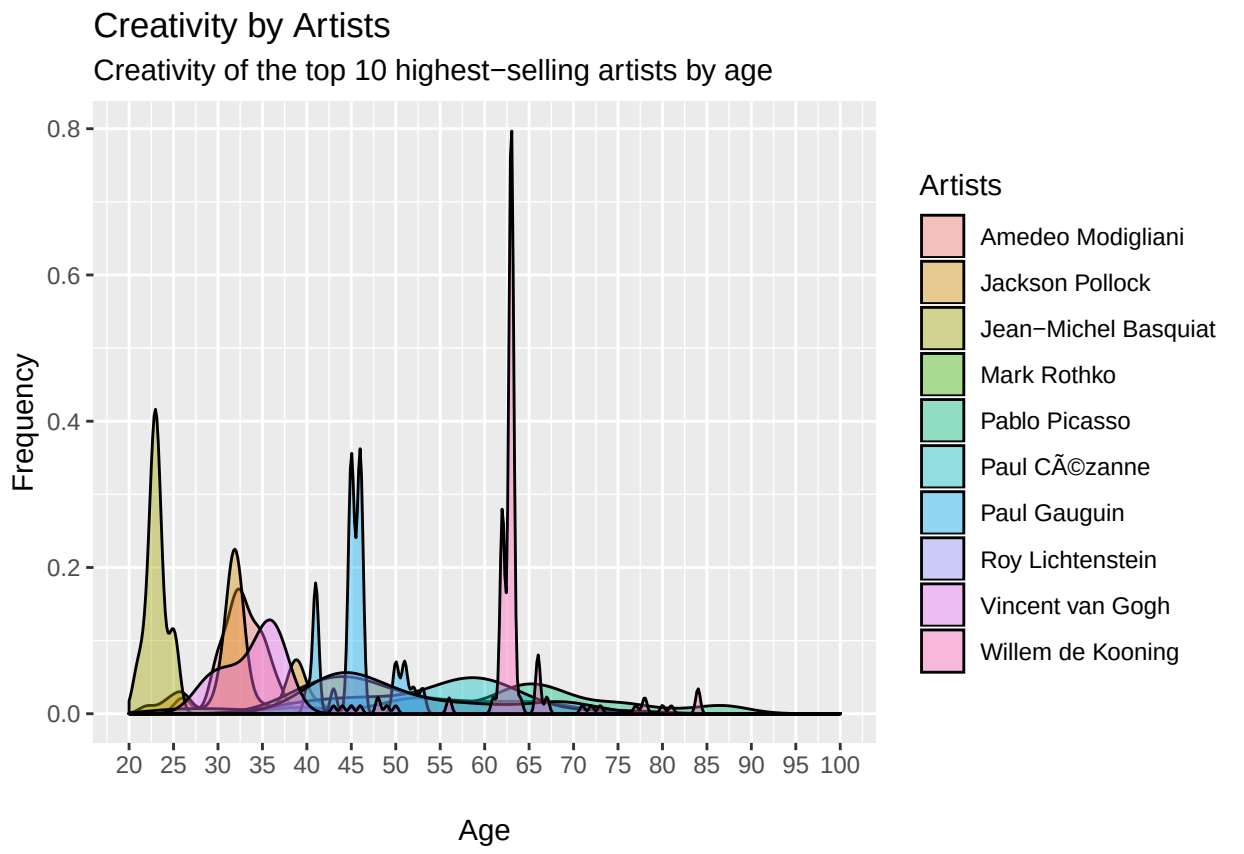
Similar analyses can be done with Bernard Tschumi and Pablo Picasso. The fact that the peaks line up with known periods of production for each artist suggests that the sample of artworks in the MoMA are a relatively accurate representation of an artist's work.

```
Artists=art$Artist
Artists = data.frame(Artists)
Artists$Age = art$Age
Artists=melt(Artists)
Selected=Artists[Artists$Artists == 'Pablo Picasso' |Artists$Artists == 'Bernard Tschumi'|Artists$Artis
ggplot(Selected, aes(x = value,fill= Artists,y=..density..)) +
  geom_density(alpha=0.2)+
  scale_x_continuous(breaks = seq(0, 100, 5), lim = c(20, 100))+
  labs(x="\nAge",y = "Frequency")+
  ggtitle("Creativity by Artists",
          subtitle = "Creativity of specific artists by age")
```



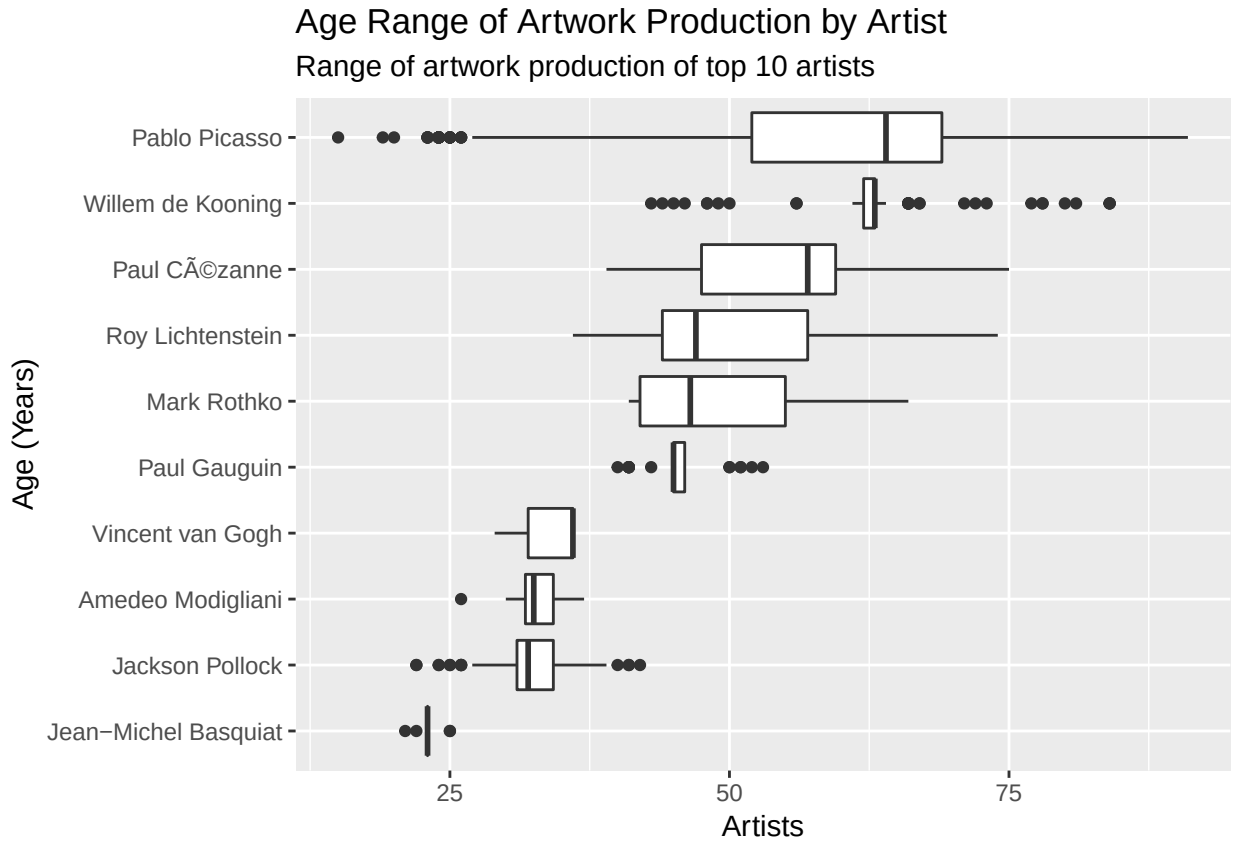
We can also look at the artists whose works are among the 11 most expensive paintings in history. A quick glance at their superimposed density curves for the amount of work that they created within the MoMA by age shows that for the most part, their patterns of production are very different. This suggests that the artist's age when creating a piece does not play a factor in determining the value of a piece.

```
richest = c('Mark Rothko', 'Roy Lichtenstein', 'Jean-Michel Basquiat', 'Amedeo Modigliani', 'Pablo Pica
ten_artists = filter(Artists, Artists %in% richest)
ggplot(ten_artists, aes(x = value,fill= Artists,y=..density..)) +
  geom_density(alpha=0.4)+
  scale_x_continuous(breaks = seq(0, 100, 5), lim = c(20, 100))+
  labs(x="\nAge",y = "Frequency")+
  ggtitle("Creativity by Artists",
          subtitle = "Creativity of the top 10 highest-selling artists by age")
```

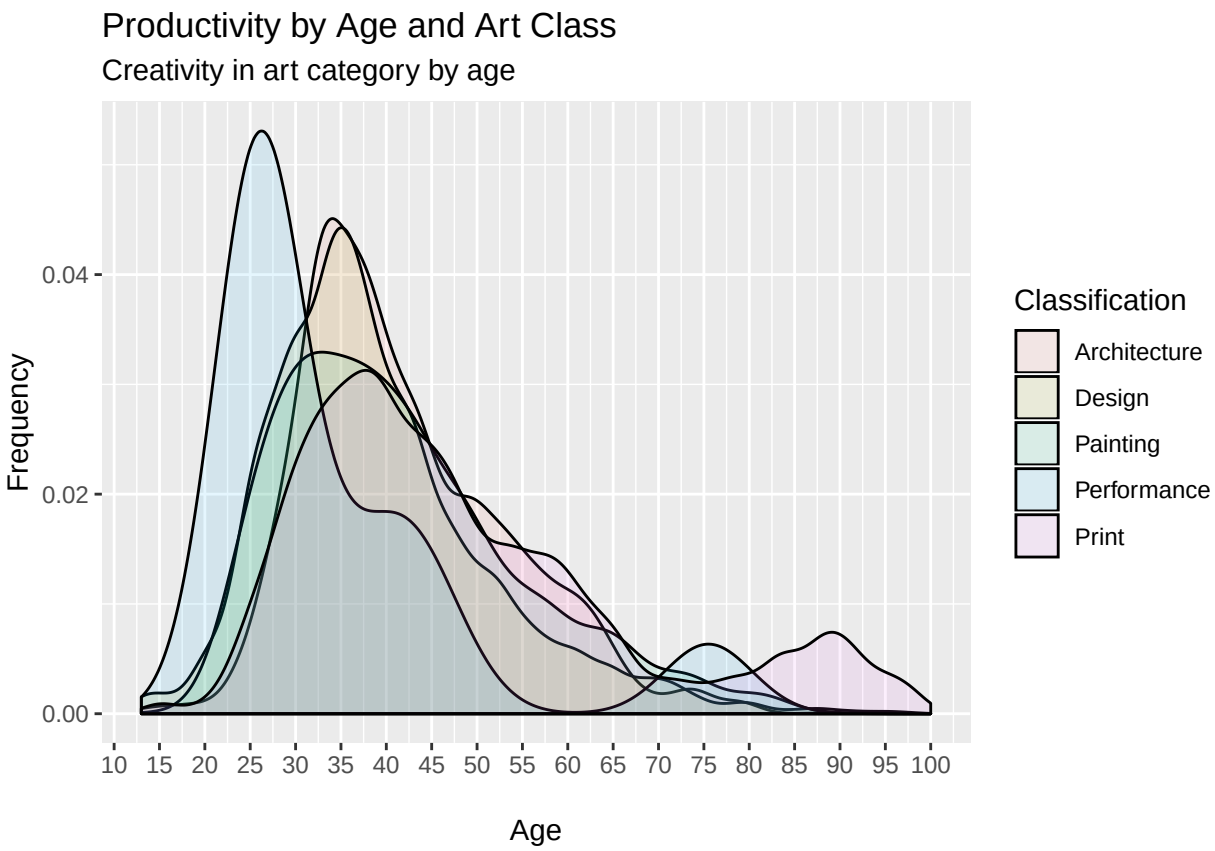



```
clean_art_ages <- read.csv(file="MOMA_art_cleaned_ages.csv")
#Usual cleaned data, just added ages column to or?g?nal cleaned one
```

However, it is difficult to properly compare the artists with the above plot, as some of them overlap and the colors are difficult to distinguish for certain artists. The following box plots give a better sense of the timespan that each of the ten artists were creating artworks. We can see that Picasso was prolific throughout his lifetime, whereas Basquiat had an extremely short career. Cezanne and Lichtenstein had similar ranges, while de Kooning was highly active for a short period, which is why the box is so small and his range contains many outliers. Van Gogh, Modigliani, and Pollock had similar interquartile ranges as well, indicating that 50% of their work was produced during that time. We performed various analyses on this data, trying to visualize it in different ways including faceted histograms and bar plots of the proportion of artwork purchased after an artist's death. We really wanted to find out whether these artists had any similarities in terms of how old they were when they were producing their art, as that would be an indication of a possible variable contributing to the value of the art. However, we found that there was no significant common overlap in this respect between all 10 artists.

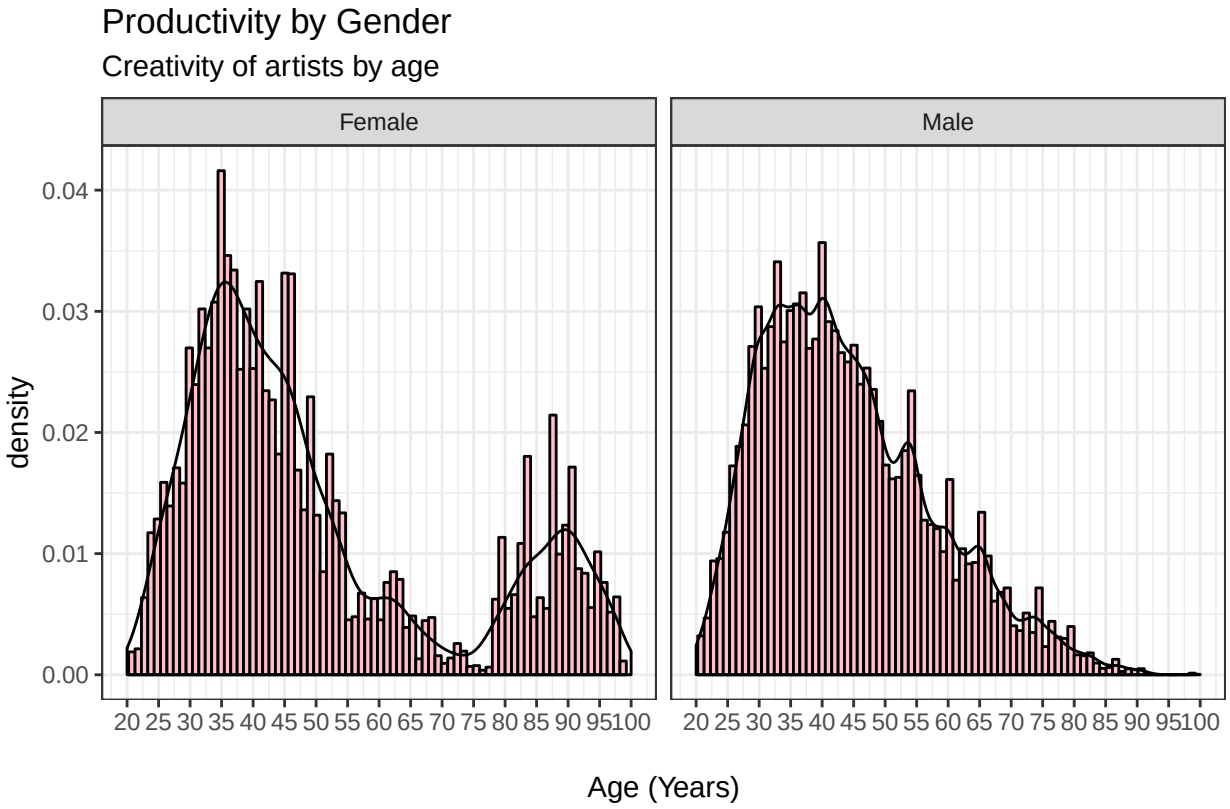


Now let's look at the effect of age on creativity with respect to art type. It is very interesting to see that **Performance Art** results in higher rates of works at a younger age, more than any other art type.



Productivity with respect to gender

We wanted to look at whether gender affected productivity levels. To do so, we created a histogram faceted on gender that shows their respective productivities. It is interesting to see that women gain productivity again after 80 years old while men steadily decrease in terms of productivity.



Type of art by gender

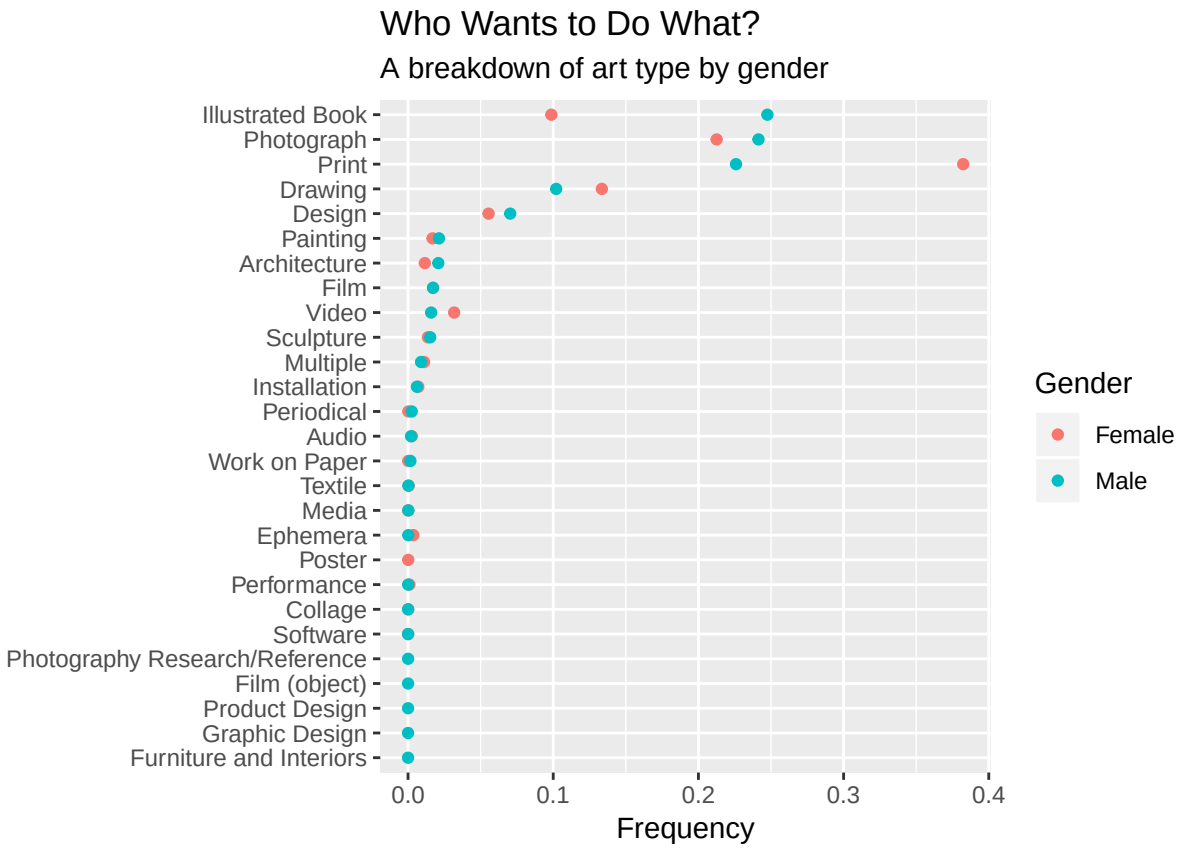
We also thought it would be interesting to break down the type of art by gender. We looked at the gender proportion within each classification by creating a Cleveland dot plot.

To construct the Cleveland dot plot, we filtered some of the variables in a sub-dataframe. There is a difference between the numbers of female and male painters. Thus, to understand the affect of gender on the selection of art class we need to normalize data. This can be achieved by grouping every gender/class combination and dividing by the total people belong to that gender group.

In the below Cleveland Dot plot, we can observe that 40% of women lean towards a painting style called print. Also known as printmaking, this type of art is defined as "The process of making artworks by printing, normally on paper. Printmaking normally covers only the process of creating prints that have an element of originality, rather than just being a photographic reproduction of a painting." on Wikipedia.

Also, when we look at male artists, we observe that illustrated books composed more than any other type of art with a ratio of 25%. In the following sections we are going to see that this art class is popular among modern French artists.

We also see that variance between the art class of male and female artists increases with as we move on y-axis.

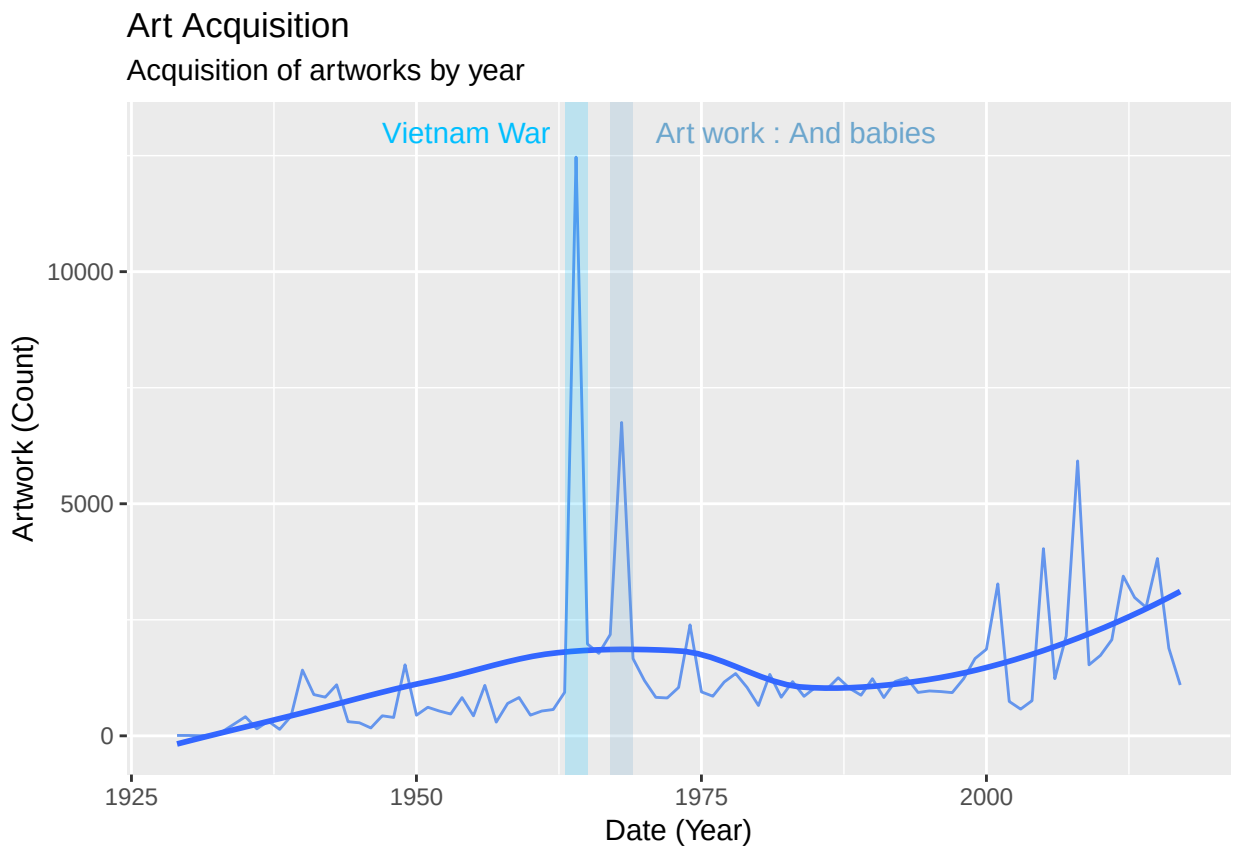


Acquisition of artwork

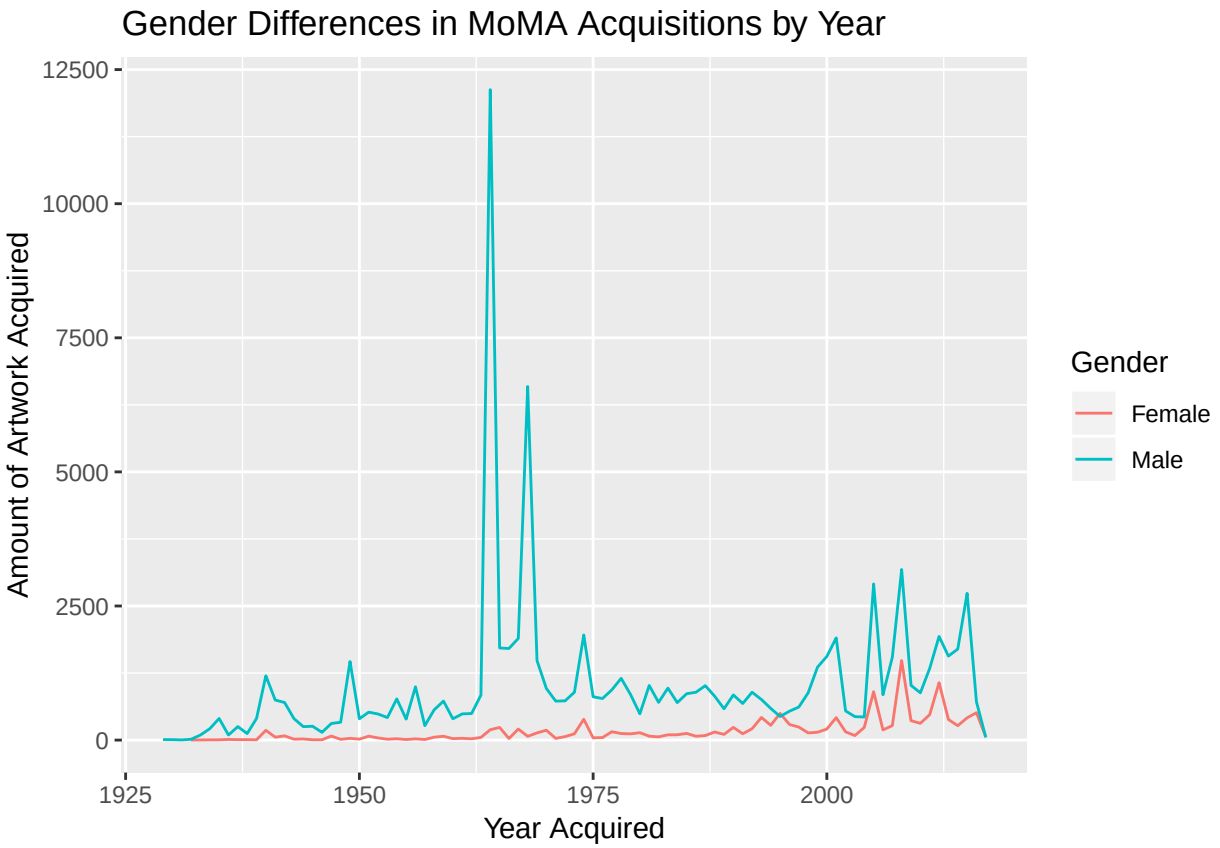
Through this plot, we see that there is a peak around the 1960s. If we look closely, the year where the peak happens corresponds to the 1964 Vietnam War. The date has been marked with a blue pointer. It is a well known fact that world problems and issues can be catalysts for art. We can thus hypothesize that the war drove interest in art and acquisitions at the museum. In this plot we can see this trend. While the modern art that has been joined to museum was 938 on 1963. It is more than 10 times on 1964 with 12468 artworks. We also see another peak during 1968-1969. That is the year the most casualties happened. Between those years, the Army of the Republic of Vietnam lost nearly 50,000 and the United States Army lost nearly 30,000 soldiers. It is also important to mention a famous artwork of MoMA during those days, called *"And babies"*, which caused a lot of controversy at the time. It was a propoganda poster against the Vietnam War which the MoMA pulled financing for at the last minute.

We also noted that many of the prints done by Eugene Atget were acquired in 1968, contributing to the second peak.

It is also interesting to note that during the Cold War, the CIA reportedly encouraged the exhibition and production of modern art. Several high ranking people at the MoMA were affiliated with the CIA, and there were several foundations secretly funded by the CIA that helped artists who were part of the Abstract Expressionism movement gain prominence by funding and facilitating exhibitions. We can see a slight rise in the general trend of acqusition during that period.



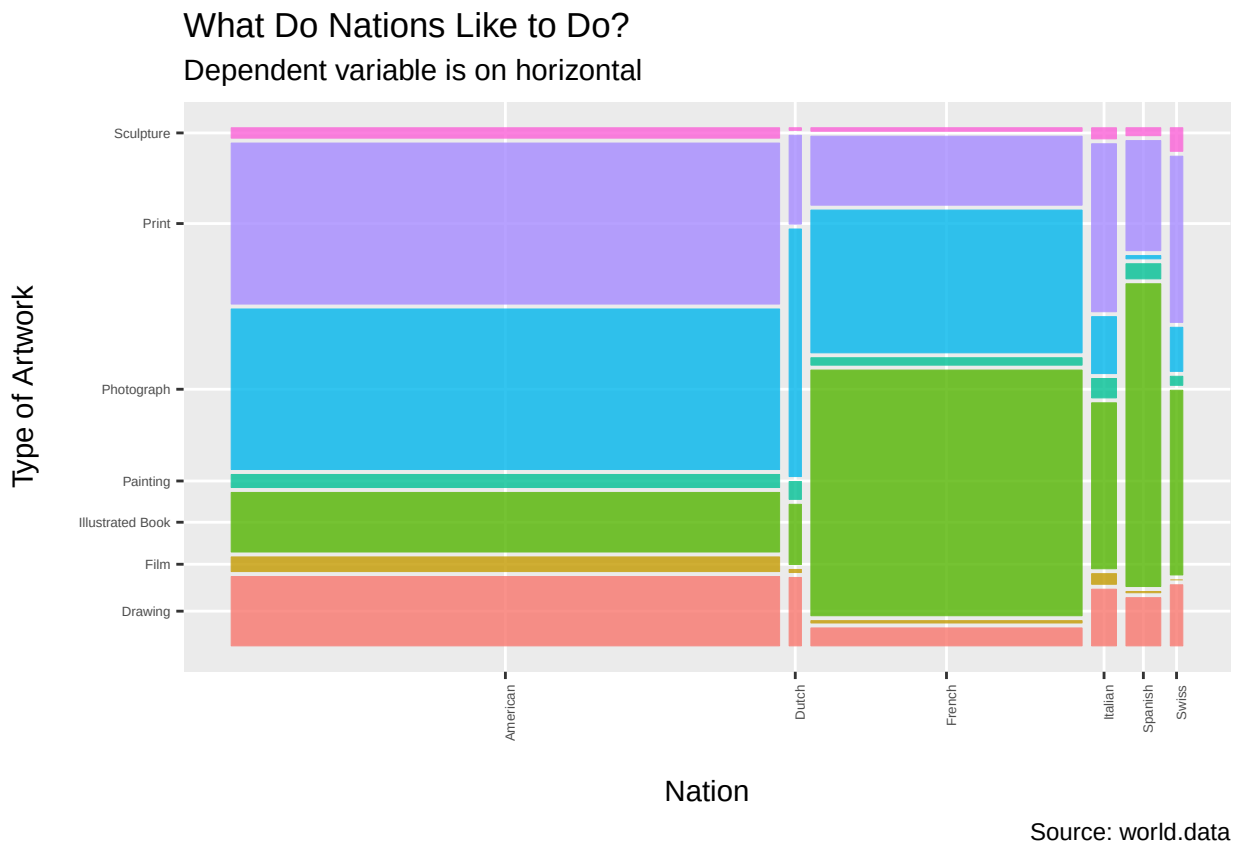
We also broke down acquisition by gender. There is no clear pattern, but unsurprisingly, the number of artworks acquired that were created by females was far less than the number of males. However, both had similar peaks in the 2000s, indicating that there was an outside factor affecting the acquisitions at that point.



Which nationality expresses themselves with what type of art?

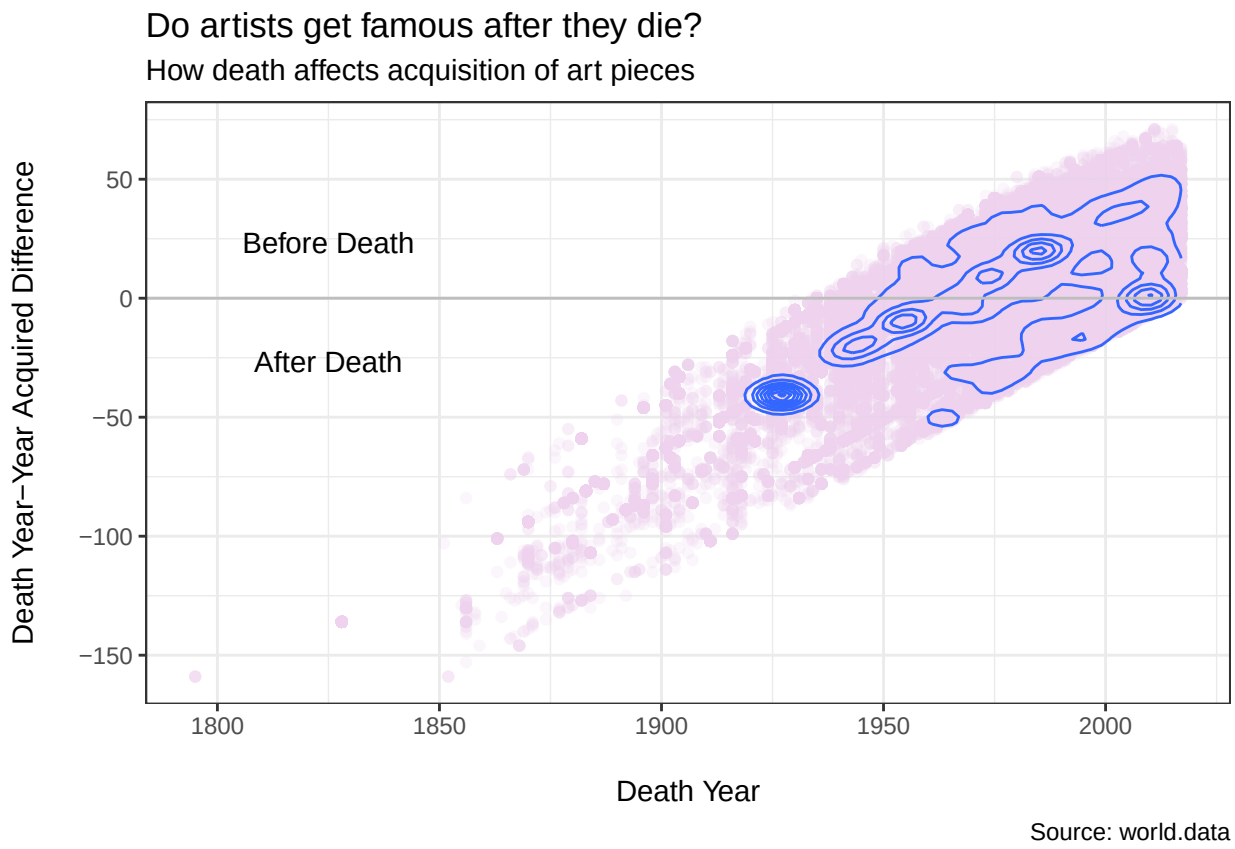
In this EDA, we are going to select 6 countries and 7 art types. The countries we selected are as follows:
1.America
2.Germany
3.French
4.Italy
5.Spain
6.Switzerland
The artists who are citizens of those countries are refered by their nationality at the below mosiac plot.
The art categories we are going to work on are:
1.Sculpture
2.Print
3.Photograph
4.Painting
5.Illustrated Book
6.Film
7.Drawing

In the below plot, we can observe that Spanish people tend to express themselves with paintings while Dutch people like to do so by Photographs and French people leaned towards illustrated books.The distribution in United states is approximately uniform and tends towards Photographs and Print.



Were artworks acquired before the artists died or after?

Generally it is a common question to ask if the demand for an artwork increases after the death of artist. In this section we are going to investigate this question. If we relate demand with fame, it would be entertaining to call our plot "Do artists get famous after they die?" To observe this variable we are going to look at the time difference between the death of artist and when the piece was acquired. If this difference is negative, than it means that the artwork has been acquired after the death of the artist. If the difference is positive, then it means that the piece is acquired before the artist has passed away.
In the below plot we see that a substantial amount of artwork has been acquired after the artist passed away. We also see that a highest density level isoplots lie under the cut-off.
At this point, one might see that there are some artworks whose artist passed away on late 18th and early 19th century. I found that interesting since modern art covers the period between 1860s to the 1970s. The Museum of Modern Art shows the predecessors of modern art and how art has evolved.



By which channel was the art acquired?

One interesting aspect of the data was the "Credit Line" column, which detailed how the art was acquired by the museum. After looking through the data and sorting by frequency, we saw that many pieces were gifts by various donors. We parsed through this column to classify each method of acquisition into a larger category, then created plots to visualize this data. Overwhelmingly, gifts comprised the largerst proportion of each department of art.

Purchases were a very small proportion of each department as well.

As a group, we discussed whether to do this plot as a Mosaic plot or a heat map. There were two reasons we chose to continue with the heat map. The first reason for the heat map was that it was less dense and less packed, thus providing a visually better data explanation. The second reason is we also wanted to provide diversity in our project as much as we could since we already had a mosaic plot.

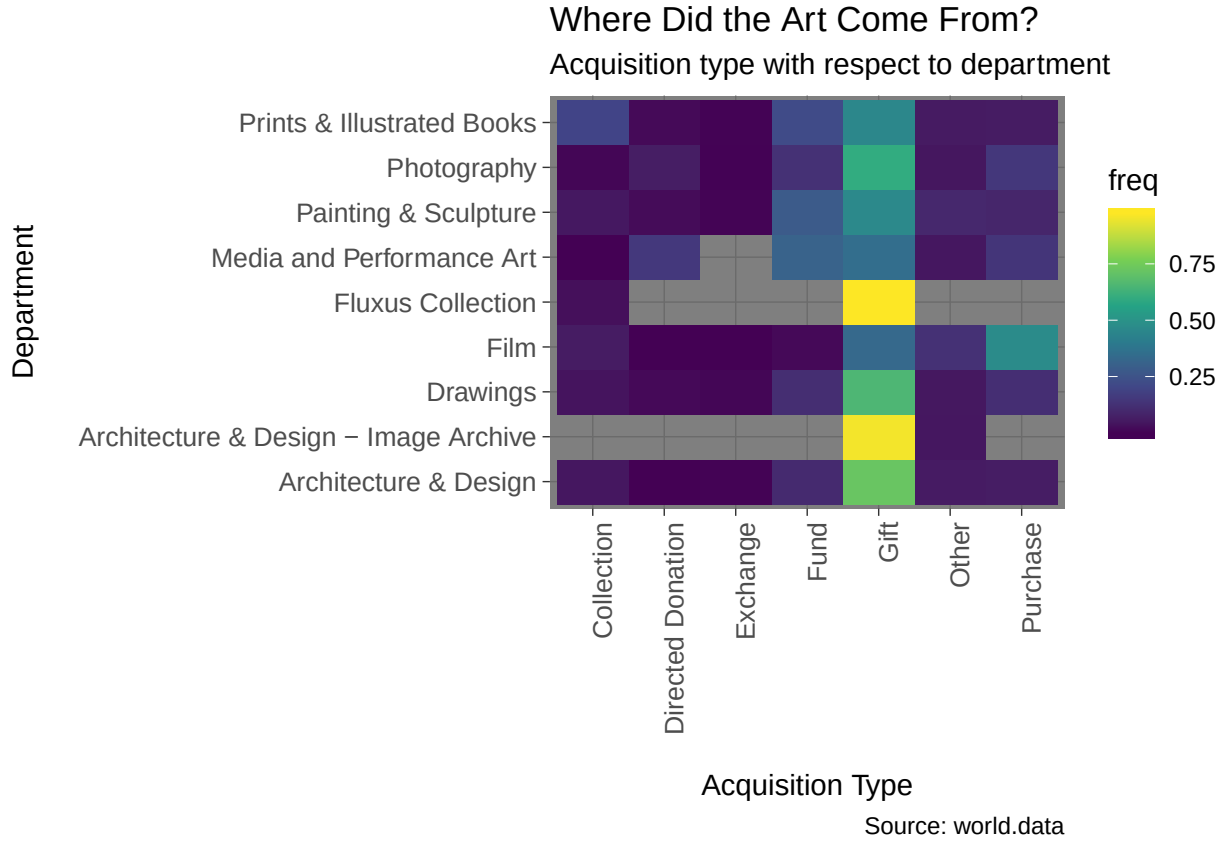
We decided to use percentage notation in this graph. That is the percentage of artworks in different departments with respect to how they were acquired. The colors corresponding to the numbers on legend adds up to one for each row.

In this graph we see that Museum of Modern arts obtains most of its artworks as gift. It can be seen that gift column has the brightest color. That corresponds to highest frequency for each department. There is a really small ratio of purchases. The purchase occurs most for film department. Thus, we can hypothesize that the budget of Museum of Modern Arts is spent mostly for Film department.

```
#art$DateAcquired <- art1$DateAcquired

art1$method <- ifelse((grepl('Purchase', art1$CreditLine) | grepl('purchase', art1$CreditLine)), "Purcha
```

```
method = art1$method
method = data.frame(method)
method$Department = art1$Department
counts2 = method %>%
  group_by(Department, method) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n))
counts2 = counts2[complete.cases(counts2), ]
ggplot(counts2, aes(method, Department)) + geom_raster(aes(fill = freq))+scale_fill_viridis_c()+theme_da
labs(x="\nAcquisition Type",y="Department\n", caption = "Source: world.data")+
  ggtitle("Where Did the Art Come From?",
          subtitle = "Acquisition type with respect to department") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1,size = 10))+
  theme(axis.text.y = element_text(hjust = 1,size = 10))
```



Geographic Patterns

In order to get a visual idea of which countries the pieces in the museum came from, we created a geographical map detailing the amount of art based on country of origin. As we can see below, the United States has the highest amount of art.

We based this template of this map on code we found online describing oil production by country (see references).

Initially, we had experimented with various ways of showing this data. We started by sorting nationalities by frequency and creating histograms, but we thought it would be visually interesting to include a map, and had

wanted at first to make this an interactive map where users could see the change in country distribution over the years. However, we weren't able to do this in time, and decided to focus on other aspects of the analysis.

This map is a good starting point to get a visual idea of the major countries and areas contributing to the artwork in the MoMA. We also created a choropleth map, but thought this one would be most striking visually and would be more accurate with respect to the color scale. We wanted to make adjustments to the color scale on the map below as well, to really find an accurate way of representing the data in color, which we could have worked on given more time.

In order to accomplish this, we used a csv file that we found online which mapped demonyms to country names (https://t2a.io/blog/normalising-nationalities-via-a-good-iso-3166-country-list/) and used Python to clean and format the nationalities that were in our original data set to their country names. We then were able to try mapping the data both using choroplethr and using map_data and the theme we found online to create a map.

This distribution is interesting overall. Though the USA is the most prevalent country in present day, modern art was introduced to the US in 1913 through an international exhibition. It would be interesting, given time, to visualize the change in distribution over the years.

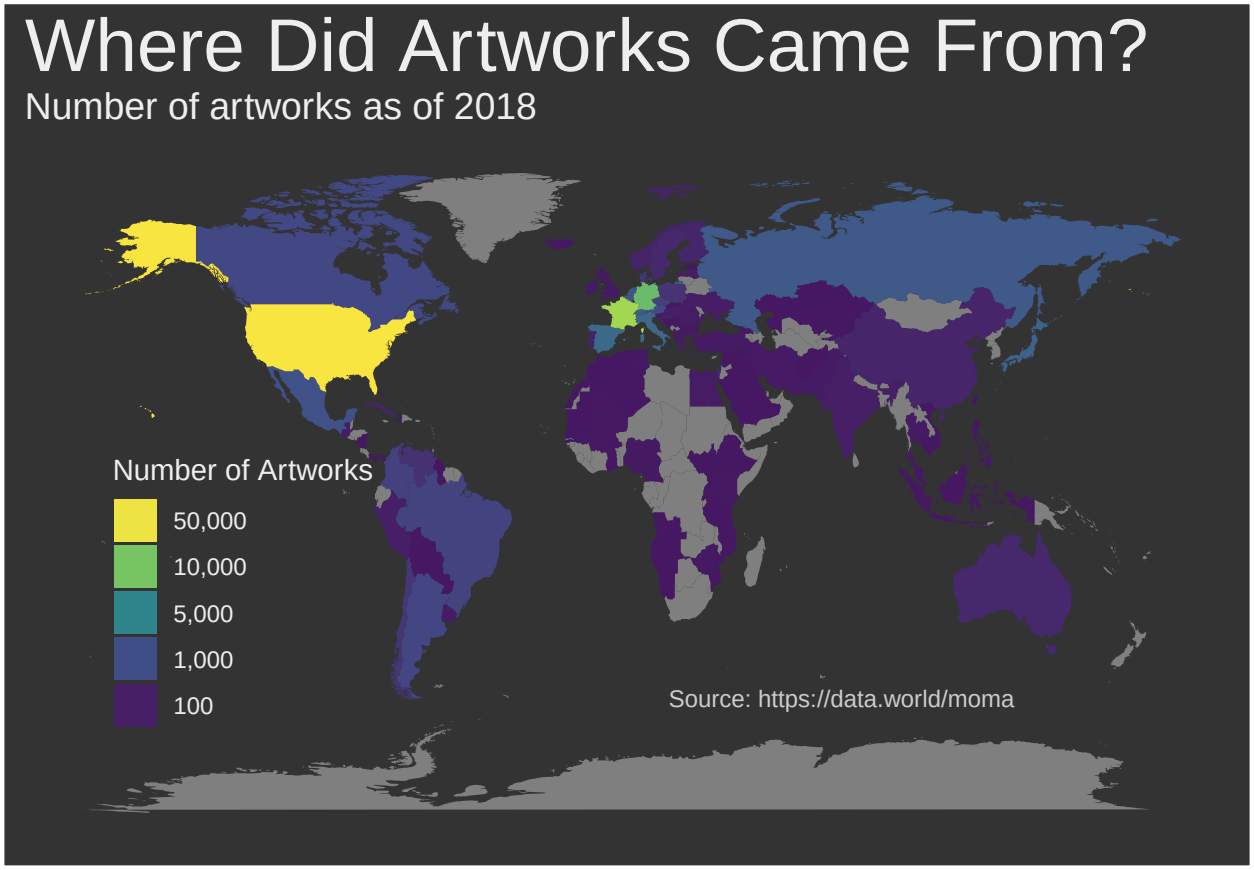
```
#detach("package:plyr", unload=TRUE)
regions <- read.csv(file="MOMA_cleaned3")
r <- regions %>% group_by(region1) %>% summarise(Freq=n()) %>% arrange(desc(Freq))
colnames(r) = c("region", "value")

map.world <- map_data('world')
mapping <- left_join( map.world, r, by = c('region' = 'region'))

theme_map <- theme(text = element_text(color = '#EEEEEE')
        ,plot.title = element_text(size = 28)
        ,plot.subtitle = element_text(size = 14)
        ,axis.ticks = element_blank()
        ,axis.text = element_blank()
        ,panel.grid = element_blank()
        ,panel.background = element_rect(fill = '#333333')
        ,plot.background = element_rect(fill = '#333333')
        ,legend.position = c(.18,.36)
        ,legend.background = element_blank()
        ,legend.key = element_blank()
        )

ggplot(mapping, aes( x = long, y = lat, group = group )) +
  geom_polygon(aes(fill = value)) +
  scale_fill_gradientn(colours = c('#461863','#404E88','#2A8A8C','#7FD157','#F9E53F')
                       ,values = scales::rescale(c(100,1000,5000,10000,50000))
                       ,labels = comma
                       ,breaks = c(100,1000,5000,10000,50000)
                       ) +
  guides(fill = guide_legend(reverse = T)) +
  labs(fill = 'Number of Artworks'
       ,title = 'Where Did Artworks Came From?'
       ,subtitle = 'Number of artworks as of 2018'
       ,x = NULL
       ,y = NULL) +
  theme_map +
  annotate(geom = 'text'
           ,label = 'Source: https://data.world/moma'
```

```
          ,x = 18, y = -55
          ,size = 3
          ,color = '#CCCCCC'
          ,hjust = 'left'
          )
```



5. Interactive Plot

Please visit https://edavart.shinyapps.io/shinyapp2/ to view our interactive output.

For our interactive portion of the project, we wanted to look at a comparison of artist productivity, the types of work they created at certain ages, and when the artworks were acquired by MoMA. We had some general questions going into this research, like if certain types of art were acquired later by MoMA because they weren't considered as valuable at the time? We also were curious about when the art was acquired by MoMA because it is commonly said that artist is only appreciated after he "dies". Is this cultural statement true? We set out to answer this and many other questions.

One early roadblock we ran into in this analysis was that there are over 13,000 artists represented in the dataset, so it would not make sense to look at every artist. As a result, we made a judgment call and chose to look at thirteen prominent artists from a variety of backgrounds that we were interested in.

With our inputs and outputs decided on, we proceeded to use the Shiny package from R, which allows us to interactively use a drop-down menu to choose one of our thirteen favorite artists. The interactivity does not stop there though. Once the points are plotted on the scatterplot, where they are colored by which medium the artist used, the user has two options either to click a point or brush over a group of points. If the user clicks a point, they will be able to see the actual artwork that was completed! This beautiful use of interactivity allows the user to see how an artist's work has evolved over time. For example, we see

that Vincent van Gogh earlier on worked on prints, but later in his artistic life he completed three paintings. Interestingly, van Gogh's most famous painting "The Starry Night" was the earliest acquired by MoMA. While it is considered a masterpiece, one may question if its prominence at MoMA so much earlier than his other two paintings (The Olive Trees and Portrait of Joseph Roulin) has led to be van Gogh's most famous work. If the user chooses to brush over instead, they will find a sortable table of the artworks that the artist completed, where they can do further analysis and indulge the data nerd in them.

Some interesting findings from this portion of the project include that many of the chosen artists (Martin, Kusama, Rothko, and Basquiat) started out their earlier years with a heavy concentration of work being drawings, but towards the end of their art lives the drawings are few and far between. Jackson Pollock is a notable exception, with the bulk of his earlier work being prints, while about half of his works at the end being drawings. While it is not totally clear why, we can find some hints that in dealing with alcoholism about midway through his career, Pollock entered an 18-month period of psychoanalysis with Dr. Joseph Henderson, who encouraged drawing as part of his treatment (https://www.csmonitor.com/1985/1024/lpoll.html). While many of his drawings in MoMA came after this point, this time period could have sparked Pollock to create more drawings in his later life.

This analysis is just a brief look at the some of the insights that can be gleaned from the interactive plots. We encourage you to play around and learn more about these great artists.

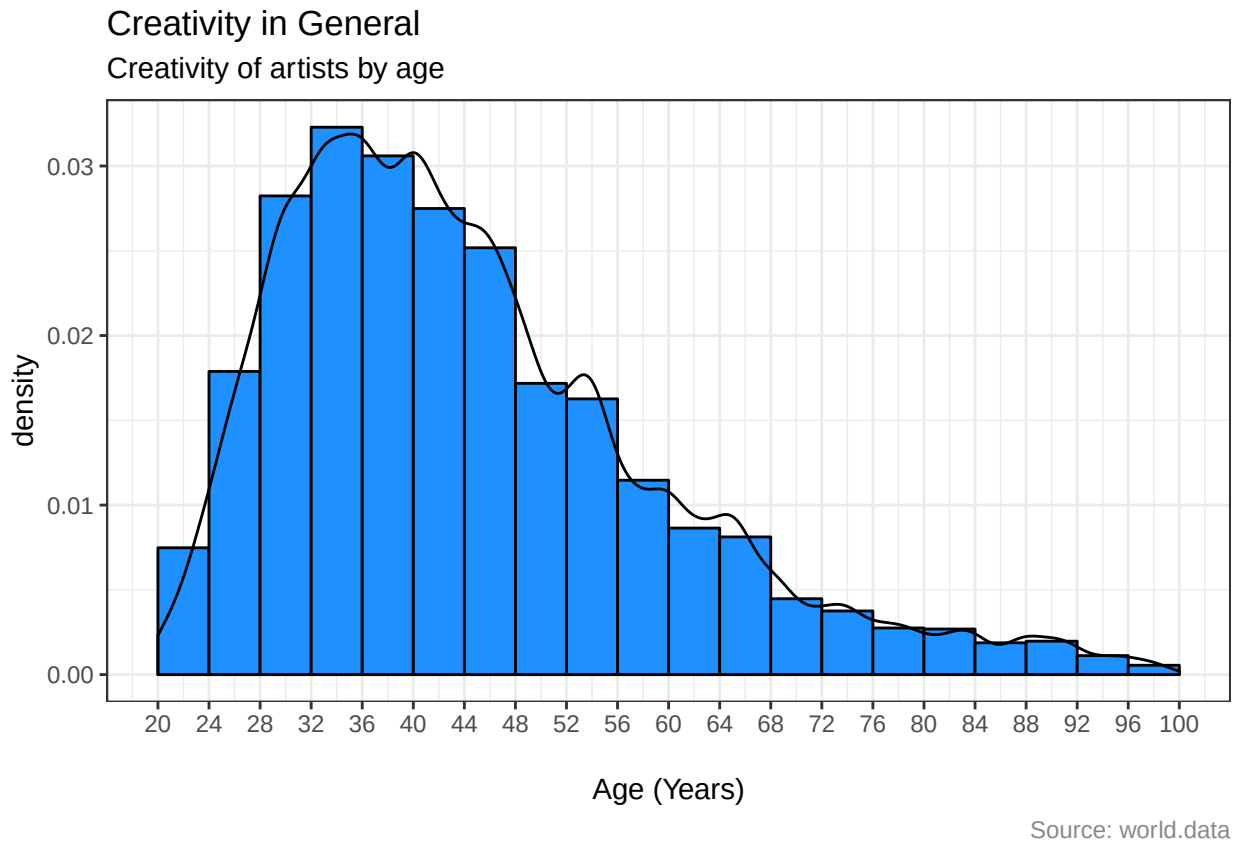
6. Executive Summary

A note preceding the summary: We have checked the graphs that are going to be presented with the color oracle software suggested in class. The colors are distinguishable and don't cause any problems in terms of interpretation of the plots. Using +scale_fill_colorblind() also did not result in any changes.

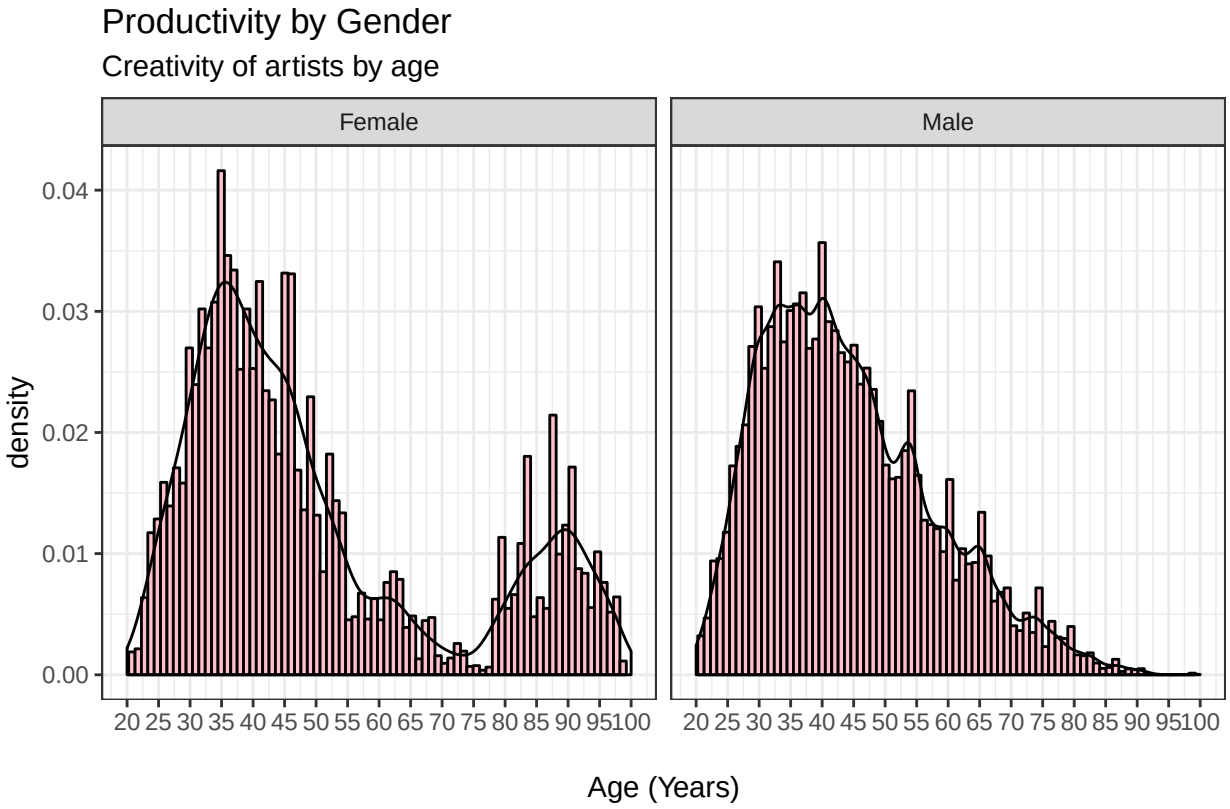
Working with the Museum of Modern Art dataset, we were able to glean many insights about the type of art and artists that are exhibited at the museum. The data is publicly available, both at https://data.world/moma, which is a platform where institutions publish their own data, and the Museum of Modern Art's Github page (https://github.com/MuseumofModernArt/collection).

After the data-cleaning process, we wanted to find out what the most productive age of artists was in general. We measured this by counting the age that an artist was at the time of creation of each artwork, and visualized it in a histogram. We then normalized it to more clearly see what the shape of this distribution was. Thus, density on the y axis is simply the count divided by the bin width so that.

It is easy to see that the range of 32-42 is the most productive times of an artist. That is the time they peak. This corresponds to known literature about the creative peaks of artists. Yet, we also can see some outliers in real life. Jean-Michel Basquiat is a good example for this situation. In his 20s, he was exhibiting his neo-expressionist paintings in galleries and museums internationally.

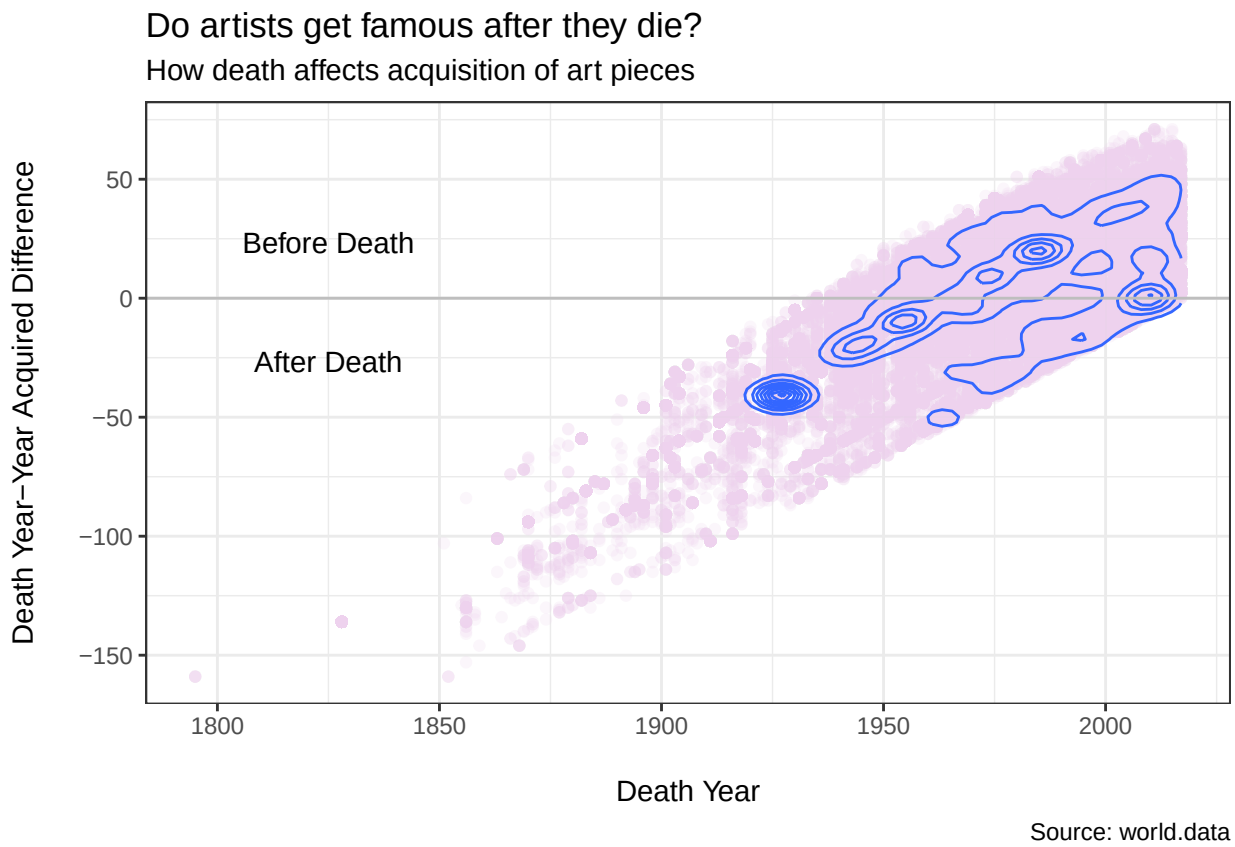


It is also interesting to think about productivity in terms of gender. It can be observed that productivity steadly decreases among male artists. However, females can reach a certain productivity level after the age of 80. Louise Bourgeois, for example, an experimental sculptor, made her greatest works and received the Lifetime Achievement in Contemporary Sculpture Award at the age of 80.



Whether art is more in demand after the artist dies is also a popular question.

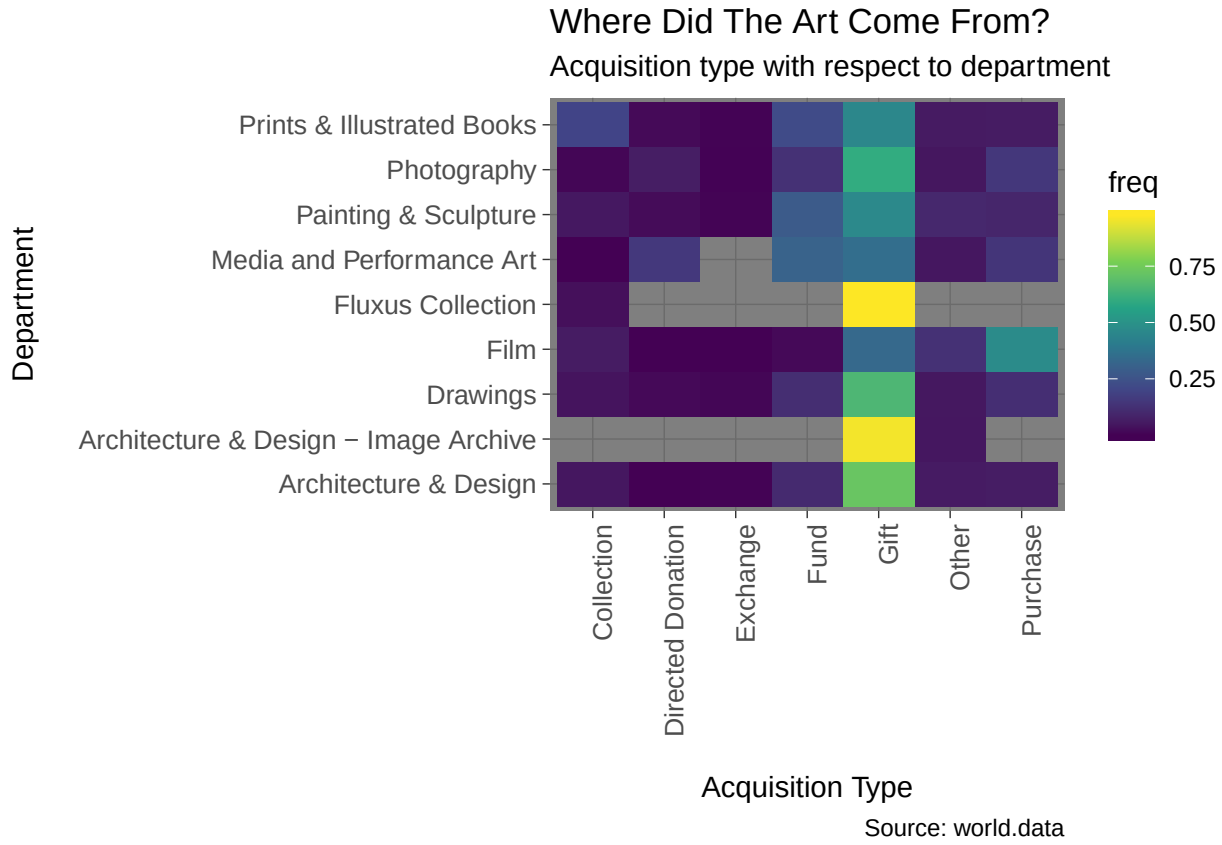
Initially, we created a scatterplot of the death year and the difference between the year the artist died and the year the piece was acquired. The contour lines on the graph indicate the areas where there are the most points. As shown below, acquiring art after approximately 50 years after artist passes away is the most common pattern, as that point contains the densest contour lines. In general, a large portion of the artwork in the MoMA was acquired after the artist's death.



Among other insights we found were that the largest proportions of acquisitions in each department were in fact gifts, and not purchased by the museum. This is something that would be interesting to explore further. The museum really does not buy most of the artwork within its departments, which is extremely interesting. At a high level, we can hypothesize that it is very prestigious to be exhibited at the MoMA, so the museum does not need to pay for the artwork it shows.

Gifts included gifts from the artist, from family members, and from private collections or individuals. There were also several directed donations, in which the museum was able to acquire a piece of art "through the generosity of" a certain patron, which were not counted as gifts in this analysis, but could be construed as such if we wanted to widen the categories.

The Film department has the highest proportion of purchases compared to the other departments. A lot of art also comes from private collections or funds, which presumably lend their pieces to the museum for viewing.



7. Conclusion

In conclusion, we had a very interesting experience performing our EDAV project on the Museum of Modern Art's data collection. We visualized many interesting data and were able to see trends regarding productivity of artists with respect to their age, productivity with respect to age and gender, and the preference of art with respect to gender.

As a team, we had lots of discussions regarding which exploratory analysis technique to use on a given data set. We tried to use as different plots as possible without losing the information that could be extracted from the data. We discovered a pattern for creativity, we investigated the preference of artworks with respect to gender, we saw how the life events of different artists affect their performance, we saw that different type of arts can be popular for different nations.

Some of our findings are also discoveries of active research fields. Yet, what may be more fascinating is the multitude of questions that came up throughout our research. Why is there a jump in art produced in this year? Why do women have a jump in art productivity in later years? Why does an artist change his or her preferred medium over time? These questions are just scratching the surface of what we can learn from the data.

Future directions for this project could involve diving deeper into more specific subsets of this data. For example, analyzing just paintings, or looking into what a specific donor has contributed to the museum. We could also take a step back and look at data from other museums, comparing and contrasting trends within them.

If we had more time on this project, we would have liked to beautify the Shiny app by adding more interactive features and plots to it. In addition, we would like to investigate the proportion of each medium used by the most productive artists. Other steps we had considered but were unable to proceed with due to lack of

data include analyzing the monetary value of the artworks in MoMA with respect to many of the features we considered in this report. Finally, bringing in other outside data such as the GDP of the nation at the time an artist created his or her work could provide value in further examination.

References

Unwin, A. (2015). Graphical data analysis with R. Boca Raton: Chapman and Hall.

Franses, P. H. (2016). When Did Classic Composers Make Their Best Work? Creativity Research Journal, 28(2), 219-221. doi:10.1080/10400419.2016.1162489

Sooke, A. (2016, October 04). Culture - Was modern art a weapon of the CIA? Retrieved from http://www.bbc.com/culture/story/20161004-was-modern-art-a-weapon-of-the-cia

Beale, C. (n.d.). The world's 10 most valuable artworks. Retrieved from https://www.weforum.org/agenda/2017/11/leonardo-da-vinci-most-expensive-artworks/

Mapping oil production by country using R. (2017, December 13). Retrieved from https://www.sharpsightlabs.com/blog/map-oil-production-country-r/