Most Expensive Painting

WebScraping, Cleaning and Plotting

Litta Jose Thottam, 135546

2024-07-02

Contents

Most Expensive Paintings

The dataset we're working with is derived from the Wikipedia page "List of most expensive paintings." This page lists the highest known prices paid for paintings, adjusted for inflation to 2023 values. The record holder is Leonardo da Vinci's "Salvator Mundi," which was sold for approximately $450.3 million in November 2017 at Christie's in New York City.

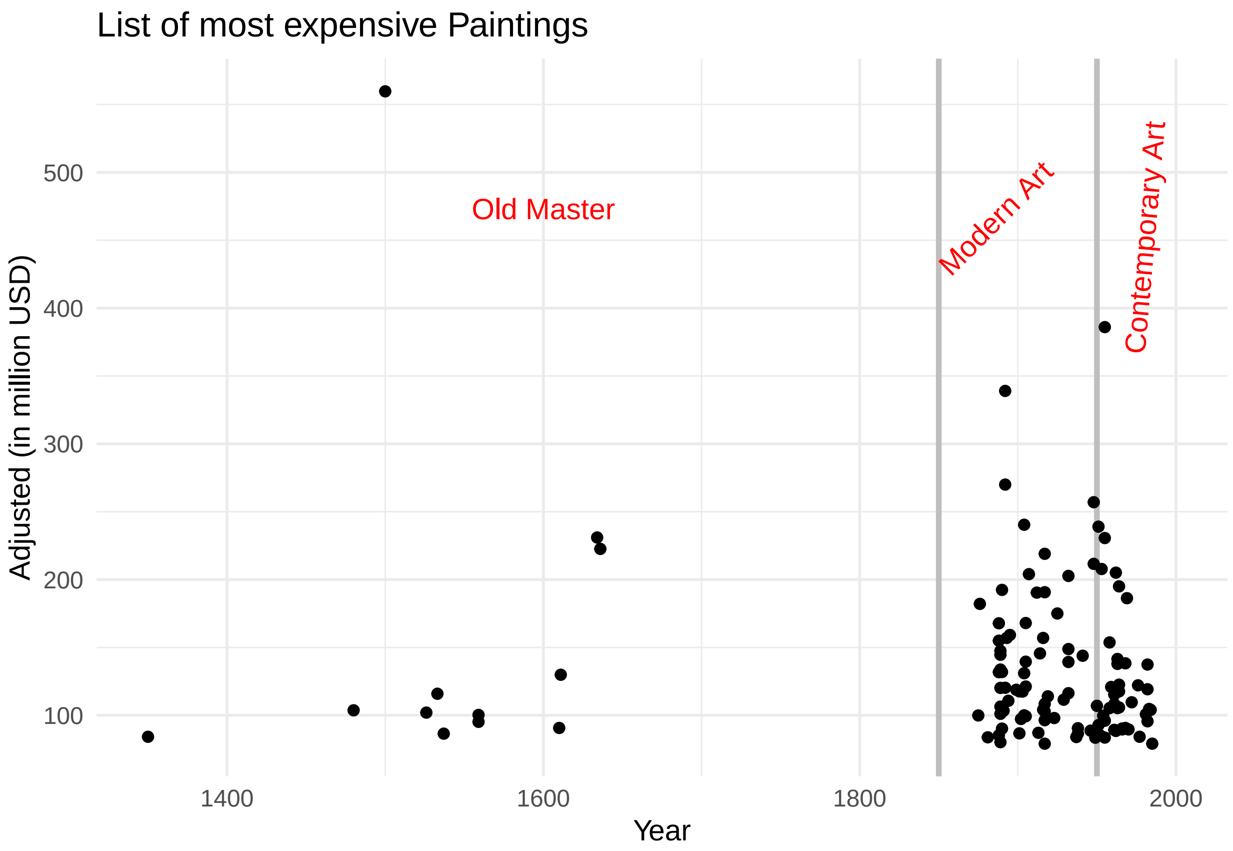
Webscrapping

```
# URL of the Wikipedia page
url <- "https://en.wikipedia.org/wiki/List_of_most_expensive_paintings"

# Read the HTML content
webpage <- read_html(url)

# Extract the table
table <- webpage %>% html_node("table.wikitable") %>% html_table(fill = TRUE)

# View the table
head(table)
## # A tibble: 6 x 11
##   `Adjusted (in million USD)` Original (in million~1 Painting Image Artist Year
##   <chr>                       <chr>                  <chr>    <lgl> <chr>  <chr>
## 1 $559.7                      $450.3                 Salvato~ NA    Leona~ c. 1~
## 2 ~$386                       ~$300                  Interch~ NA    Wille~ 1955
## 3 $339 +                      $250 +[note 3]         The Car~ NA    Paul ~ 1892~
## 4 $270                            $210[25]               Nafea F~ NA    Paul ~ 1892
## 5 ~$257                           ~$200                  Number ~ NA    Jacks~ 1948
## 6 $240.4                          $183.8                 Wassers~ NA    Gusta~ 1904~
## # i abbreviated name: 1: `Original (in million USD)`
## # i 5 more variables: `Date of sale` <chr>, `Rankat sale` <int>, Seller <chr>,
## #   Buyer <chr>, `Auction house` <chr>

# Print column names to ensure we are referencing the correct ones
colnames(table)
##  [1] "Adjusted (in million USD)" "Original (in million USD)"
##  [3] "Painting"                  "Image"
##  [5] "Artist"                    "Year"
##  [7] "Date of sale"              "Rankat sale"
##  [9] "Seller"                    "Buyer"
## [11] "Auction house"
# Clean the 'Year' column
table$Year <- gsub("\\[.*\\]", "", table$Year)
table$Year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", table$Year))

# Clean the 'Adjusted (in million USD)' column
table$`Adjusted (in million USD)` <- gsub("\\[.*\\]", "", table$`Adjusted (in million USD)`)
table$`Adjusted (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Adjusted (in million USD)`))

# Clean the 'Original (in million USD)' column
table$`Original (in million USD)` <- gsub("\\[.*\\]", "", table$`Original (in million USD)`)
table$`Original (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Original (in million USD)`))

# Remove rows with NA values in crucial columns
table <- table %>% filter(!is.na(Year), !is.na(`Adjusted (in million USD)`))

# View cleaned data
head(table)
## # A tibble: 6 x 11
##   `Adjusted (in million USD)` Original (in million~1 Painting Image Artist  Year
##                         <dbl>                  <dbl> <chr>    <lgl> <chr>  <dbl>
## 1                        560.                   450. Salvato~ NA    Leona~  1500
## 2                        386                    300  Interch~ NA    Wille~  1955
## 3                        339                    250  The Car~ NA    Paul ~  1892
## 4                        270                    210  Nafea F~ NA    Paul ~  1892
## 5                        257                    200  Number ~ NA    Jacks~  1948
## 6                        240.                   184. Wassers~ NA    Gusta~  1904
## # i abbreviated name: 1: `Original (in million USD)`
## # i 5 more variables: `Date of sale` <chr>, `Rankat sale` <int>, Seller <chr>,
## #   Buyer <chr>, `Auction house` <chr>
```

Relationship between Creation Time and Prices (10 pts)

The goal of the graph is to visualize the relationship between the creation year of the paintings and their adjusted auction prices, providing insights into how art prices have evolved over time and highlighting different art periods.

```
ggplot(table, aes(x = Year, y = `Adjusted (in million USD)`)) +
  geom_vline(xintercept = c(1850, 1950), size = 1, color = "grey")+


  geom_point() +
  annotate("text", x = 1600, y = 450, label = "Old Master", color = "red", angle = 360, vjust = -1) +
  annotate("text", x = 1900, y = 450, label = "Modern Art", color = "red", angle = 45, vjust = -1) +
  annotate("text", x = 2000, y = 450, label = "Contemporary Art", color = "red", angle = 85, vjust = -1
  theme_minimal()+


labs(title = "List of most expensive Paintings")
```

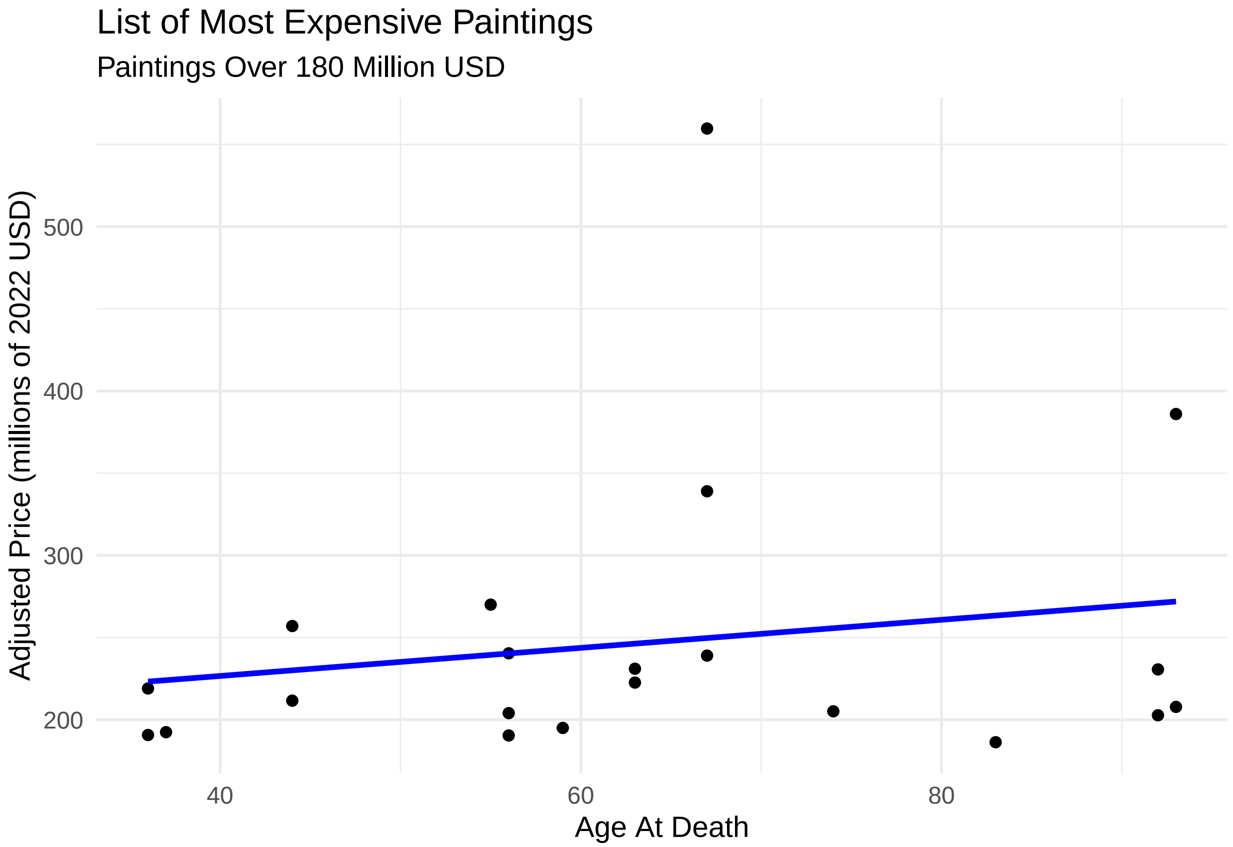


Age at Death and Prices (2 pt BONUS)

I hypothesis that the age at which an artist dies might be positively correlated with the price of their most expensive paintings. The idea is that as artists grow older, they gain more experience, hone their skills, and produce higher quality work, which in turn might fetch higher prices at auction.

As artists age, they accumulate more experience, which can translate into better technical skills, a deeper understanding of their medium, and more sophisticated artistic concepts. This accumulated expertise can result in more refined and impactful artworks, which are valued more highly by collectors and institutions. Over time, artists often evolve in their style and technique. This evolution can lead to the creation of unique and mature works that stand out in the art market. Late works by artists can often be seen as the culmination of their artistic journey, incorporating all their learning and experimentation.

```r
# Load required libraries
library(rvest)
library(dplyr)
library(ggplot2)
library(lubridate)

# URL of the Wikipedia page
url <- "https://en.wikipedia.org/wiki/List_of_most_expensive_paintings"

# Read the HTML content
webpage <- read_html(url)

# Extract the table
table <- webpage %>% html_node("table.wikitable") %>% html_table(fill = TRUE)

# Clean the 'Year' column
table$Year <- gsub("\\[.*\\]", "", table$Year)
table$Year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", table$Year))

# Clean the 'Adjusted (in million USD)' column
table$`Adjusted (in million USD)` <- gsub("\\[.*\\]", "", table$`Adjusted (in million USD)`)
table$`Adjusted (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Adjusted (in million USD)`))

# Clean the 'Original (in million USD)' column
table$`Original (in million USD)` <- gsub("\\[.*\\]", "", table$`Original (in million USD)`)
table$`Original (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Original (in million USD)`))

# Remove rows with NA values in crucial columns
table <- table %>% filter(!is.na(Year), !is.na(`Adjusted (in million USD)`))

# Extract painter's birth year (if available)
table$Artist <- gsub("\\(.*\\)", "", table$Artist)
table <- table %>% mutate(BirthYear = case_when(
  Artist == "Leonardo da Vinci"     ~ 1452,
  Artist == "Vincent van Gogh"      ~ 1853,
  Artist == "Pablo Picasso"         ~ 1881,
  Artist == "Andy Warhol"           ~ 1928,
  Artist == "Jackson Pollock"       ~ 1912,
  Artist == "Willem de Kooning"     ~ 1904,
  Artist == "Paul Cézanne"          ~ 1839,
  Artist == "Paul Gauguin"          ~ 1848,
  Artist == "Rembrandt"             ~ 1606,  # Example: Rembrandt van Rijn
  Artist == "Michelangelo"          ~ 1475,  # Example: Michelangelo Buonarroti
  Artist == "Claude Monet"          ~ 1840,  # Example: Claude Monet
  Artist == "Johannes Vermeer"      ~ 1632,  # Example: Johannes Vermeer
  Artist == "Salvador Dalí"         ~ 1904,  # Example: Salvador Dalí
  Artist == "Georgia O'Keeffe"      ~ 1887,  # Example: Georgia O'Keeffe
  Artist == "Frida Kahlo"           ~ 1907,  # Example: Frida Kahlo
  Artist == "Mark Rothko"           ~ 1903,  # Example: Mark Rothko
  Artist == "Edvard Munch"          ~ 1863,  # Example: Edvard Munch
  Artist == "Gustav Klimt"          ~ 1862,  # Example: Gustav Klimt
  Artist == "Amedeo Modigliani"     ~ 1884,  # Example: Amedeo Modigliani
  Artist == "Henri Matisse"         ~ 1869,  # Example: Henri Matisse
  Artist == "Diego Rivera"          ~ 1886,  # Example: Diego Rivera
  Artist == "Peter Paul Rubens"     ~ 1577,  # Example: Peter Paul Rubens
  Artist == "Francis Bacon"         ~ 1909,  # Example: Francis Bacon
  Artist == "Jean-Michel Basquiat"  ~ 1960,  # Example: Jean-Michel Basquiat
  Artist == "Joan Miró"             ~ 1893,  # Example: Joan Miró
  Artist == "Roy Lichtenstein"      ~ 1923,  # Example: Roy Lichtenstein
  Artist == "Édouard Manet"         ~ 1832,  # Example: Édouard Manet
  Artist == "Banksy"                ~ 1974,  # Example: Banksy
  TRUE ~ NA_real_
))

# Define the birth and death years for each artist
artists <- c("Leonardo da Vinci", "Vincent van Gogh", "Pablo Picasso", "Andy Warhol", "Jackson Pollock"
             "Rembrandt", "Michelangelo", "Claude Monet", "Johannes Vermeer", "Salvador Dalí", "Georgia
             "Edvard Munch", "Gustav Klimt", "Amedeo Modigliani", "Henri Matisse", "Diego Rivera", "Pet
             "Jean-Michel Basquiat", "Joan Miró", "Roy Lichtenstein", "Édouard Manet", "Banksy")
birth_years <- c(1452, 1853, 1881, 1928, 1912, 1904, 1839, 1848,
                 1606, 1475, 1840, 1632, 1904, 1887, 1907, 1903,
                 1863, 1862, 1884, 1869, 1886, 1577, 1909,
                 1960, 1893, 1923, 1832, 1974)
death_years <- c(1519, 1890, 1973, 1987, 1956, 1997, 1906, 1903,
                 1669, 1564, 1926, 1675, 1989, 1986, 1954, 1970,
                 1944, 1918, 1920, 1954, 1957, 1640, 1992,
                 1988, 1983, 1997, 1883, 0)

# Calculate age at death
age_at_death <- death_years - birth_years

# Merge age at death data with the table
table_with_age <- merge(table, data.frame(Artist = artists, AgeAtDeath = age_at_death), by = "Artist", 

# Filter out rows with missing age at death
table_with_age <- table_with_age %>%
  filter(!is.na(AgeAtDeath))

# Filter out rows with adjusted price less than 180 million
table_with_age <- table_with_age %>%
  filter(`Adjusted (in million USD)` > 180)

# Plot age at death vs adjusted price
ggplot(table_with_age, aes(x = AgeAtDeath, y = `Adjusted (in million USD)`)) +
  geom_point(color = "black") + # Set color to black for all points
  geom_smooth(method = "lm", color = "blue", fill = "grey", se = FALSE) + # Add a linear model with a g
  labs(
    title = "List of Most Expensive Paintings",
     subtitle = "Paintings Over 180 Million USD",
    x = "Age At Death",
    y = "Adjusted Price (millions of 2022 USD)"
  ) +
  theme_minimal()
```



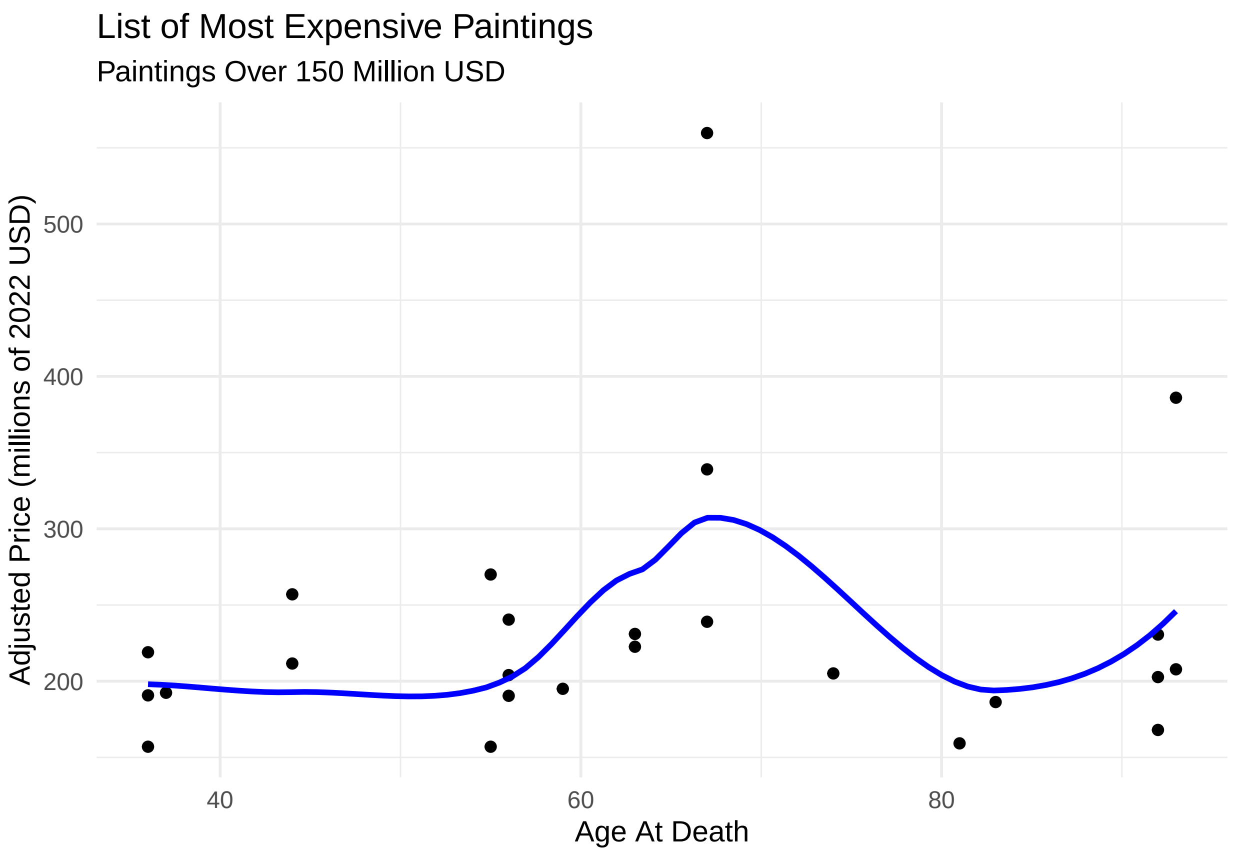
Age at Painting and Prices (2 pt BONUS)

Artists might reach a peak in their creative and technical abilities around mid-life (around 50 years old). This period could be when they produce their most significant and valuable works. This peak period could be a result of a combination of accumulated experience and creativity, leading to masterpieces that are highly valued in the art market. Art created during an artist's peak years might be perceived as the most representative of their mature style and technical mastery. Collectors and investors may place higher value on these works.

```r
# Load required libraries
library(rvest)
library(dplyr)
library(ggplot2)
library(lubridate)

# URL of the Wikipedia page
url <- "https://en.wikipedia.org/wiki/List_of_most_expensive_paintings"

# Read the HTML content
webpage <- read_html(url)

# Extract the table
table <- webpage %>% html_node("table.wikitable") %>% html_table(fill = TRUE)

# Clean the 'Year' column
```

```r
table$Year <- gsub("\\[.*\\]", "", table$Year)
table$Year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", table$Year))

# Clean the 'Adjusted (in million USD)' column
table$`Adjusted (in million USD)` <- gsub("\\[.*\\]", "", table$`Adjusted (in million USD)`)
table$`Adjusted (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Adjusted (in million USD)`))

# Clean the 'Original (in million USD)' column
table$`Original (in million USD)` <- gsub("\\[.*\\]", "", table$`Original (in million USD)`)
table$`Original (in million USD)` <- as.numeric(gsub("[^0-9.]", "", table$`Original (in million USD)`))

# Remove rows with NA values in crucial columns
table <- table %>% filter(!is.na(Year), !is.na(`Adjusted (in million USD)`))

# Extract painter's birth year (if available)
table$Artist <- gsub("\\(.*\\)", "", table$Artist)
table <- table %>% mutate(BirthYear = case_when(
  Artist == "Leonardo da Vinci"     ~ 1452,
  Artist == "Vincent van Gogh"      ~ 1853,
  Artist == "Pablo Picasso"         ~ 1881,
  Artist == "Andy Warhol"           ~ 1928,
  Artist == "Jackson Pollock"       ~ 1912,
  Artist == "Willem de Kooning"     ~ 1904,
  Artist == "Paul Cézanne"          ~ 1839,
  Artist == "Paul Gauguin"          ~ 1848,
  Artist == "Rembrandt"             ~ 1606,
  Artist == "Michelangelo"          ~ 1475,
  Artist == "Claude Monet"          ~ 1840,
  Artist == "Johannes Vermeer"      ~ 1632,
  Artist == "Salvador Dalí"         ~ 1904,
  Artist == "Georgia O'Keeffe"      ~ 1887,
  Artist == "Frida Kahlo"           ~ 1907,
  Artist == "Mark Rothko"           ~ 1903,
  Artist == "Edvard Munch"          ~ 1863,
  Artist == "Gustav Klimt"          ~ 1862,
  Artist == "Amedeo Modigliani"     ~ 1884,
  Artist == "Henri Matisse"         ~ 1869,
  Artist == "Diego Rivera"          ~ 1886,
  Artist == "Peter Paul Rubens"     ~ 1577,
  Artist == "Francis Bacon"         ~ 1909,
  Artist == "Jean-Michel Basquiat"  ~ 1960,
  Artist == "Joan Miró"             ~ 1893,
  Artist == "Roy Lichtenstein"      ~ 1923,
  Artist == "Édouard Manet"         ~ 1832,
  TRUE ~ NA_real_
))

# Define the birth and death years for each artist
artists <- c("Leonardo da Vinci", "Vincent van Gogh", "Pablo Picasso", "Andy Warhol", "Jackson Pollock"
             "Rembrandt", "Michelangelo", "Claude Monet", "Johannes Vermeer", "Salvador Dalí", "Georgia
             "Edvard Munch", "Gustav Klimt", "Amedeo Modigliani", "Henri Matisse", "Diego Rivera", "Pete
             "Jean-Michel Basquiat", "Joan Miró", "Roy Lichtenstein", "Édouard Manet")
```

```r
birth_years <- c(1452, 1853, 1881, 1928, 1912, 1904, 1839, 1848,
                 1606, 1475, 1840, 1632, 1904, 1887, 1907, 1903,
                 1863, 1862, 1884, 1869, 1886, 1577, 1909,
                 1960, 1893, 1923, 1832)

death_years <- c(1519, 1890, 1973, 1987, 1956, 1997, 1906, 1903,
                 1669, 1564, 1926, 1675, 1989, 1986, 1954, 1970,
                 1944, 1918, 1920, 1954, 1957, 1640, 1992,
                 1988, 1983, 1997, 1883)

# Calculate age at death
age_at_death <- death_years - birth_years

# Merge age at death data with the table
table_with_age <- merge(table, data.frame(Artist = artists, AgeAtDeath = age_at_death), by = "Artist", a

# Filter out rows with missing age at death
table_with_age <- table_with_age %>%
  filter(!is.na(AgeAtDeath))

# Filter out rows with adjusted price less than 150 million
table_with_age <- table_with_age %>%
  filter(`Adjusted (in million USD)` > 150)

# Plot age at death vs adjusted price with loess smoothing
ggplot(table_with_age, aes(x = AgeAtDeath, y = `Adjusted (in million USD)`)) +
  geom_point(color = "black") + # Set color to black for all points
  geom_smooth(method = "loess", color = "blue", fill = "grey", se = FALSE) + # Add a loess smooth curve
  labs(
    title = "List of Most Expensive Paintings",
    subtitle = "Paintings Over 150 Million USD",
    x = "Age At Death",
    y = "Adjusted Price (millions of 2022 USD)"
  ) +
  theme_minimal()
```


List of Most Expensive Paintings
Paintings Over 150 Million USD
Adjusted Price (millions of 2022 USD)
500
400
300
200
40
60
80
Age At Death