Art History Data Anaysis

Tianyi Ma

05.09.2023

In high school I took AP Art History. Although I only got a 3 out of 5 in the exam, which is one of my worst scores, I loved the subject since I get to look at a sea of artworks. TidyTuesday has a ton of fun datasets, but this one stood out to me because they actually extracted data from two Art History textbooks, and I think one of them is actually my high school Art History textbook. I’m compelled to analyze it at this point.

# Read Dataset

library(tidytuesdayR)

## Warning: package 'tidytuesdayR' was built under R version 4.2.3

# tuesdata <- tidytuesdayR::tt\_load('2023-01-17')  
# tuesdata <- tidytuesdayR::tt\_load(2023, week = 03)  
  
# arthistory <- tuesdata$arthistory

It seems like the TidyTuesday version of the dataset only has artist.csv. This is a little disappointing since I originally considered the artwork data to be the meat of this dataset. Luckily the github README file for this dataset cited its original publish site, so I went on there and there it is, the instruction of downloading the full dataset. Let’s do it.

# install.packages("arthistory")

Now let’s take a quick look of the inside of the worksgardner dataset.

library("arthistory")

## Warning: package 'arthistory' was built under R version 4.2.3

head(worksgardner, 5)

## # A tibble: 5 × 24  
## artist\_name edition\_number title\_of\_work publication\_year page\_number\_of\_image  
## <chr> <dbl> <chr> <dbl> <chr>   
## 1 Aaron Doug… 9 Noah's Ark 1991 965   
## 2 Aaron Doug… 10 Noah's Ark 1996 1053   
## 3 Aaron Doug… 11 Noah's Ark 2001 1030   
## 4 Aaron Doug… 12 Noah's Ark 2005 990   
## 5 Aaron Doug… 13 Noah's Ark 2009 937   
## # ℹ 19 more variables: artist\_unique\_id <dbl>, artist\_nationality <chr>,  
## # artist\_gender <chr>, artist\_race <chr>, artist\_ethnicity <chr>,  
## # height\_of\_work\_in\_book <dbl>, width\_of\_work\_in\_book <dbl>,  
## # height\_of\_text <dbl>, width\_of\_text <dbl>, extra\_text\_height <dbl>,  
## # extra\_text\_width <dbl>, area\_of\_work\_in\_book <dbl>, area\_of\_text <dbl>,  
## # extra\_text\_area <dbl>, total\_area\_text <dbl>, total\_space <dbl>,  
## # page\_area <dbl>, space\_ratio\_per\_page <dbl>, book <chr>

Getting basic info on the data:

str(worksgardner)

## tibble [2,325 × 24] (S3: tbl\_df/tbl/data.frame)  
## $ artist\_name : chr [1:2325] "Aaron Douglas" "Aaron Douglas" "Aaron Douglas" "Aaron Douglas" ...  
## $ edition\_number : num [1:2325] 9 10 11 12 13 14 15 16 14 15 ...  
## $ title\_of\_work : chr [1:2325] "Noah's Ark" "Noah's Ark" "Noah's Ark" "Noah's Ark" ...  
## $ publication\_year : num [1:2325] 1991 1996 2001 2005 2009 ...  
## $ page\_number\_of\_image : chr [1:2325] "965" "1053" "1030" "990" ...  
## $ artist\_unique\_id : num [1:2325] 1 1 1 1 1 1 1 1 2 2 ...  
## $ artist\_nationality : chr [1:2325] "American" "American" "American" "American" ...  
## $ artist\_gender : chr [1:2325] "Male" "Male" "Male" "Male" ...  
## $ artist\_race : chr [1:2325] "Black or African American" "Black or African American" "Black or African American" "Black or African American" ...  
## $ artist\_ethnicity : chr [1:2325] "Not Hispanic or Latinx" "Not Hispanic or Latinx" "Not Hispanic or Latinx" "Not Hispanic or Latinx" ...  
## $ height\_of\_work\_in\_book: num [1:2325] 11.3 12.1 12.3 12.3 12.8 12.8 12.7 7.9 14 12.8 ...  
## $ width\_of\_work\_in\_book : num [1:2325] 8.5 8.9 8.8 8.8 9.3 9.3 9.2 19 10.2 9.2 ...  
## $ height\_of\_text : num [1:2325] 14.5 12.4 10.8 15.7 15 18.8 21.2 14.7 4.5 16.2 ...  
## $ width\_of\_text : num [1:2325] 8.4 9 9 8.9 9.3 9.3 9.2 13.9 9.3 9.2 ...  
## $ extra\_text\_height : num [1:2325] 0 0 0 0 0 0 0 0 9.2 0 ...  
## $ extra\_text\_width : num [1:2325] 0 0 0 0 0 0 0 0 8.8 0 ...  
## $ area\_of\_work\_in\_book : num [1:2325] 96 108 108 108 119 ...  
## $ area\_of\_text : num [1:2325] 121.8 111.6 97.2 139.7 139.5 ...  
## $ extra\_text\_area : num [1:2325] 0 0 0 0 0 ...  
## $ total\_area\_text : num [1:2325] 121.8 111.6 97.2 139.7 139.5 ...  
## $ total\_space : num [1:2325] 218 219 205 248 259 ...  
## $ page\_area : num [1:2325] 616 586 677 658 649 ...  
## $ space\_ratio\_per\_page : num [1:2325] 0.353 0.374 0.303 0.377 0.398 ...  
## $ book : chr [1:2325] "gardner" "gardner" "gardner" "gardner" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ArtistName = col\_character(),  
## .. EditionNumber = col\_double(),  
## .. TitleofWork = col\_character(),  
## .. Year = col\_double(),  
## .. PageNumberofImage = col\_character(),  
## .. ArtistUniqueID = col\_double(),  
## .. ArtistNationality = col\_character(),  
## .. ArtistGender = col\_character(),  
## .. ArtistRace = col\_character(),  
## .. ArtistEthnicity = col\_character(),  
## .. `HeightofWorkinGardner(cm)\*` = col\_double(),  
## .. `WidthofWorkinGardner(cm)` = col\_double(),  
## .. `LengthofText(cm)` = col\_double(),  
## .. `WidthofText(cm)` = col\_double(),  
## .. `ExtraTextLength(cm)` = col\_double(),  
## .. `ExtraTextWidth(cm)` = col\_double(),  
## .. `AreaofWorkinGardner(cm^2)` = col\_double(),  
## .. `AreaofText(cm^2)` = col\_double(),  
## .. `ExtraTextArea(cm^2)` = col\_double(),  
## .. `TotalAreaText(cm^2)` = col\_double(),  
## .. `TotalSpace(cm^2)` = col\_double(),  
## .. `PageArea(cm^2)` = col\_double(),  
## .. SpaceRatioPerPage = col\_double(),  
## .. Book = col\_character(),  
## .. ...25 = col\_skip(),  
## .. ...26 = col\_skip(),  
## .. ...27 = col\_skip(),  
## .. ...28 = col\_skip(),  
## .. ...29 = col\_skip(),  
## .. ...30 = col\_skip(),  
## .. ...31 = col\_skip(),  
## .. ...32 = col\_skip(),  
## .. ...33 = col\_skip(),  
## .. ...34 = col\_skip(),  
## .. ...35 = col\_skip(),  
## .. ...36 = col\_skip(),  
## .. ...37 = col\_skip(),  
## .. ...38 = col\_skip(),  
## .. ...39 = col\_skip(),  
## .. ...40 = col\_skip(),  
## .. ...41 = col\_skip(),  
## .. ...42 = col\_skip(),  
## .. ...43 = col\_skip(),  
## .. ...44 = col\_skip(),  
## .. ...45 = col\_skip(),  
## .. ...46 = col\_skip(),  
## .. ...47 = col\_skip(),  
## .. ...48 = col\_skip(),  
## .. ...49 = col\_skip(),  
## .. ...50 = col\_skip(),  
## .. ...51 = col\_skip(),  
## .. ...52 = col\_skip()  
## .. )

That is a little disappointing – I was expecting to see descriptive texts associated with each artwork. But this is okay. The size of the artwork and its texts also could be interesting.

Right now, though, we observe that the first 5 rows are all work called “Noah’s Ark” by Aaron Douglas, just in different editions of the book. Because I don’t care about different editions and sizes right now, I’ll simply remove every edition other than the latest one for every author and their artwork. Right now, I want to analyze artist and their artwork info. Later, we can try answer questions like “how many artworks are included in each edition” or “How many artworks are added for each edition”.

The following snippet group row values by column artist\_name and title\_of\_work, then only retain the last row. Because the edition row values are in descending order, this should do the trick. Checking the first two rows shows that the code is working.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

worksgardner\_lean <- worksgardner %>%   
 group\_by(artist\_name, title\_of\_work) %>%  
 filter(row\_number() == n()) %>%  
 ungroup()  
head(worksgardner\_lean, 2)

## # A tibble: 2 × 24  
## artist\_name edition\_number title\_of\_work publication\_year page\_number\_of\_image  
## <chr> <dbl> <chr> <dbl> <chr>   
## 1 Aaron Doug… 15 Noah's Ark 2016 912   
## 2 Aaron Doug… 16 From Slavery… 2020 932   
## # ℹ 19 more variables: artist\_unique\_id <dbl>, artist\_nationality <chr>,  
## # artist\_gender <chr>, artist\_race <chr>, artist\_ethnicity <chr>,  
## # height\_of\_work\_in\_book <dbl>, width\_of\_work\_in\_book <dbl>,  
## # height\_of\_text <dbl>, width\_of\_text <dbl>, extra\_text\_height <dbl>,  
## # extra\_text\_width <dbl>, area\_of\_work\_in\_book <dbl>, area\_of\_text <dbl>,  
## # extra\_text\_area <dbl>, total\_area\_text <dbl>, total\_space <dbl>,  
## # page\_area <dbl>, space\_ratio\_per\_page <dbl>, book <chr>

Further looking into some basic statistic in the new dataframe shows that the number of data points shrunk from 2300+ to ~750 artist-artwork pairs.

str(worksgardner\_lean)

## tibble [746 × 24] (S3: tbl\_df/tbl/data.frame)  
## $ artist\_name : chr [1:746] "Aaron Douglas" "Aaron Douglas" "Adélaïde Labille-Guiard" "Adolphe William Bouguereau" ...  
## $ edition\_number : num [1:746] 15 16 16 6 13 14 16 3 4 10 ...  
## $ title\_of\_work : chr [1:746] "Noah's Ark" "From Slavery through Reconstruction, from Aspects of Negro Life" "Self-Portrait with Two Pupils" "The Birth of Venus" ...  
## $ publication\_year : num [1:746] 2016 2020 2020 1975 2009 ...  
## $ page\_number\_of\_image : chr [1:746] "912" "932" "784" "683" ...  
## $ artist\_unique\_id : num [1:746] 1 1 2 3 3 4 4 5 5 5 ...  
## $ artist\_nationality : chr [1:746] "American" "American" "French" "French" ...  
## $ artist\_gender : chr [1:746] "Male" "Male" "Female" "Male" ...  
## $ artist\_race : chr [1:746] "Black or African American" "Black or African American" "White" "White" ...  
## $ artist\_ethnicity : chr [1:746] "Not Hispanic or Latinx" "Not Hispanic or Latinx" "Not Hispanic or Latinx" "Not Hispanic or Latinx" ...  
## $ height\_of\_work\_in\_book: num [1:746] 12.7 7.9 12.8 10 13.6 8.6 11.2 10.3 8 10.7 ...  
## $ width\_of\_work\_in\_book : num [1:746] 9.2 19 9.2 7.3 9.3 14.8 19 12.3 8.5 13.5 ...  
## $ height\_of\_text : num [1:746] 21.2 14.7 24.8 9.5 6.9 12.4 5.6 7.7 8 14 ...  
## $ width\_of\_text : num [1:746] 9.2 13.9 9.3 7 9.3 9.3 9.3 5.9 6.1 9 ...  
## $ extra\_text\_height : num [1:746] 0 0 0 0 0 0 0 0 0 0 ...  
## $ extra\_text\_width : num [1:746] 0 0 0 0 0 0 0 0 0 0 ...  
## $ area\_of\_work\_in\_book : num [1:746] 117 150 118 73 126 ...  
## $ area\_of\_text : num [1:746] 195 204.3 230.6 66.5 64.2 ...  
## $ extra\_text\_area : num [1:746] 0 0 0 0 0 0 0 0 0 0 ...  
## $ total\_area\_text : num [1:746] 195 204.3 230.6 66.5 64.2 ...  
## $ total\_space : num [1:746] 312 354 348 140 191 ...  
## $ page\_area : num [1:746] 647 652 652 468 649 ...  
## $ space\_ratio\_per\_page : num [1:746] 0.482 0.544 0.534 0.298 0.294 ...  
## $ book : chr [1:746] "gardner" "gardner" "gardner" "gardner" ...

Now we can answer some question like “who’s the most popular artist according to this textbook?”. We will show this as a bar graph.

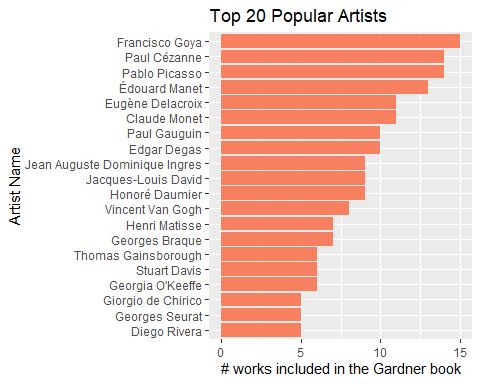
library('ggplot2')

## Warning: package 'ggplot2' was built under R version 4.2.3

library('forcats')

## Warning: package 'forcats' was built under R version 4.2.3

artist\_work\_num\_rank <- worksgardner\_lean %>%   
 count(artist\_name, sort = T)  
  
artist\_work\_num\_rank %>%  
 mutate(artist\_name = fct\_reorder(artist\_name, n)) %>%  
 slice(1:20) %>%  
 ggplot(., aes(x=artist\_name, y=n)) +   
 geom\_bar(stat = "identity", fill="#f68060") +  
 coord\_flip() +  
 ggtitle('Top 20 Popular Artists') +   
 xlab('Artist Name') +   
 ylab('# works included in the Gardner book')



TBD