HODP_R_Bootcamp

Vanessa + Jason

10/14/2020

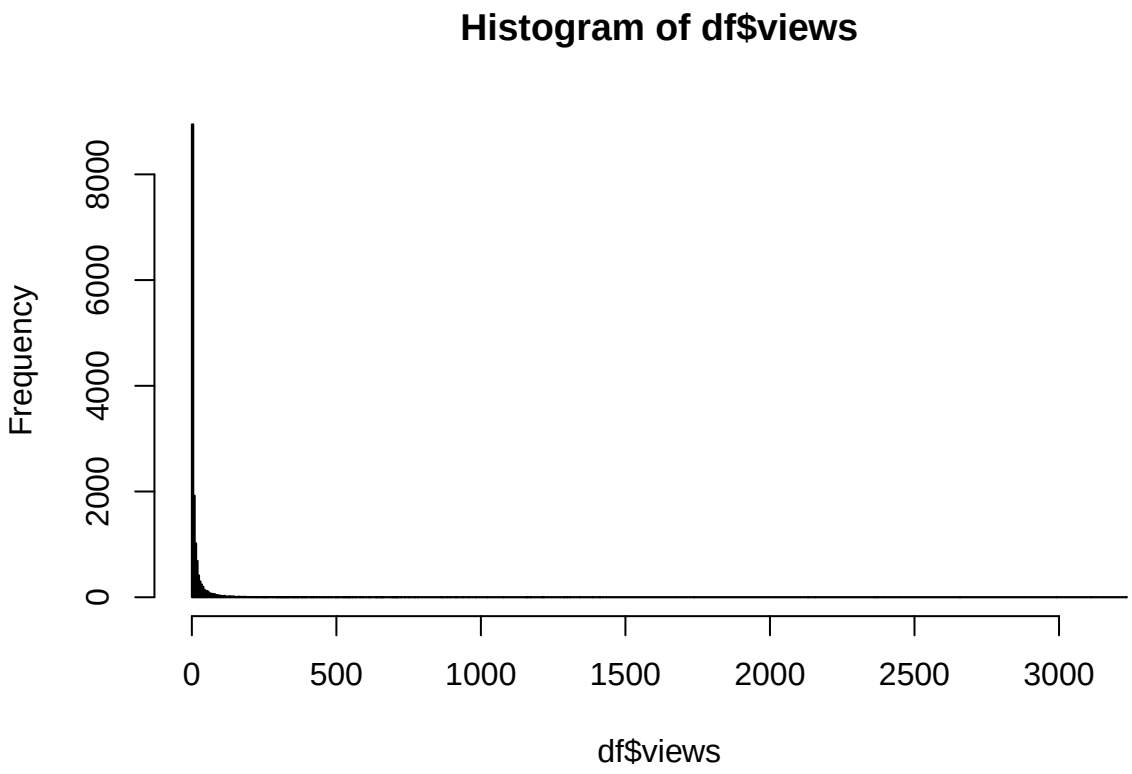
First checking the dataset as a whole

```
df <- read.csv("MuseumDataFull1015.csv")
summary(df)
```

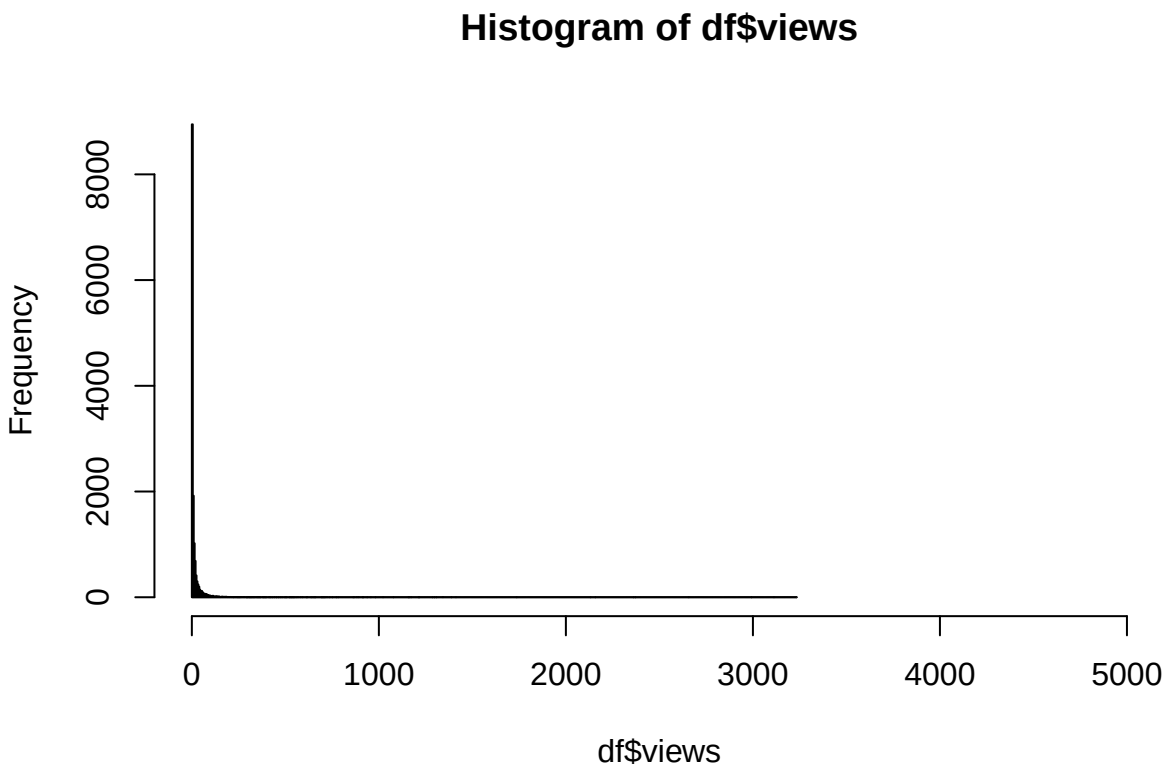
```
##     period          accessionyear     dateend         title
##  Length:15000       Min.   :1765   Min.   :-5000   Length:15000
##  Class :character   1st Qu.:1953   1st Qu.:    0   Class :character
##  Mode  :character   Median :1990   Median : 1863   Mode  :character
##                     Mean   :1981   Mean   : 1194
##                     3rd Qu.:2011   3rd Qu.: 1940
##                     Max.   :2020   Max.   : 2019
##                     NA's   :4679
##      url              division           century          objectnumber
##  Length:15000       Length:15000       Length:15000       Length:15000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  totaluniquepageviews   datebegin       culture          imagepermissionlevel
##  Min.   :   0.00      Min.   :-6000   Length:15000       Min.   :0.0000
##  1st Qu.:   1.00      1st Qu.:    0   Class :character   1st Qu.:0.0000
##  Median :   3.00      Median : 1858   Mode  :character   Median :0.0000
##  Mean   :  18.46      Mean   : 1184                      Mean   :0.2358
##  3rd Qu.:  12.00      3rd Qu.: 1938                      3rd Qu.:0.0000
##  Max.   :3234.00      Max.   : 2019                      Max.   :2.0000
##
##       rank              id          department
##  Min.   :    24   Min.   :  1415   Length:15000
##  1st Qu.: 66586   1st Qu.:138183   Class :character
##  Median :135162   Median :208706   Mode  :character
##  Mean   :135635   Mean   :199517
##  3rd Qu.:206463   3rd Qu.:272689
##  Max.   :273302   Max.   :370142
##
```

EDA of totaluniquepageviews Extremely right skewed

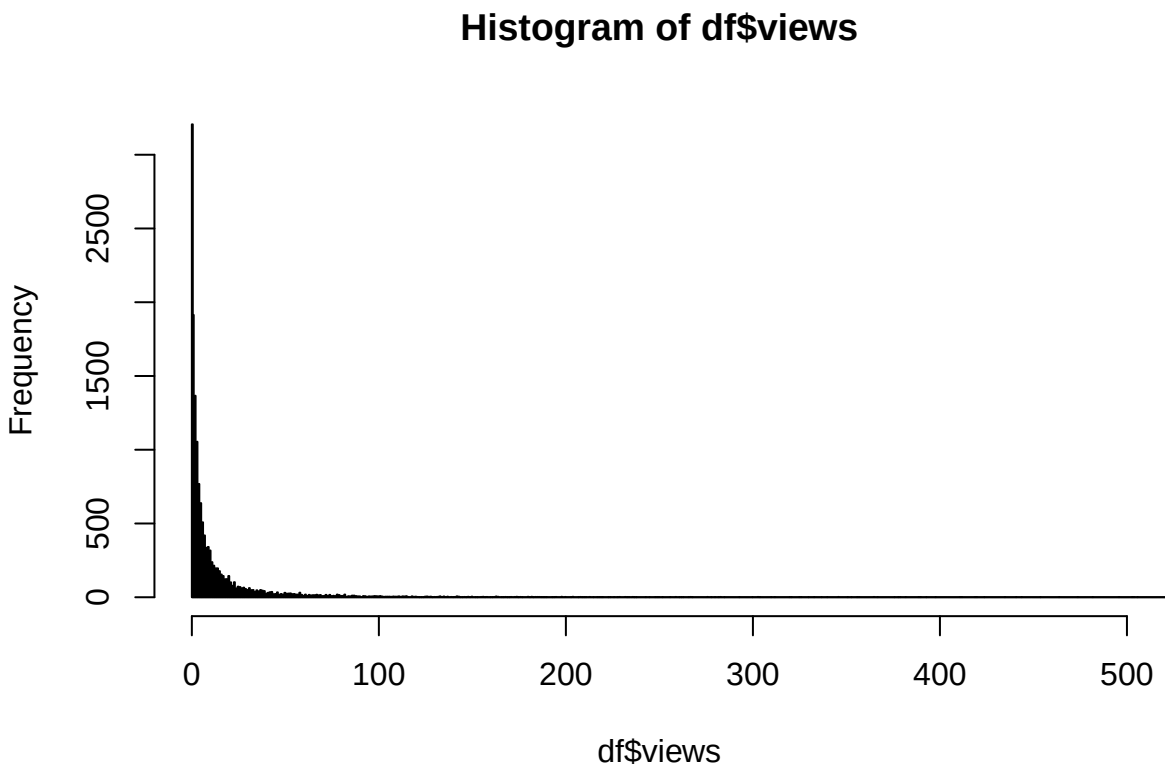
```
names(df)[names(df) == "totaluniquepageviews"] <- "views"
hist(df$views, breaks = 500)
```



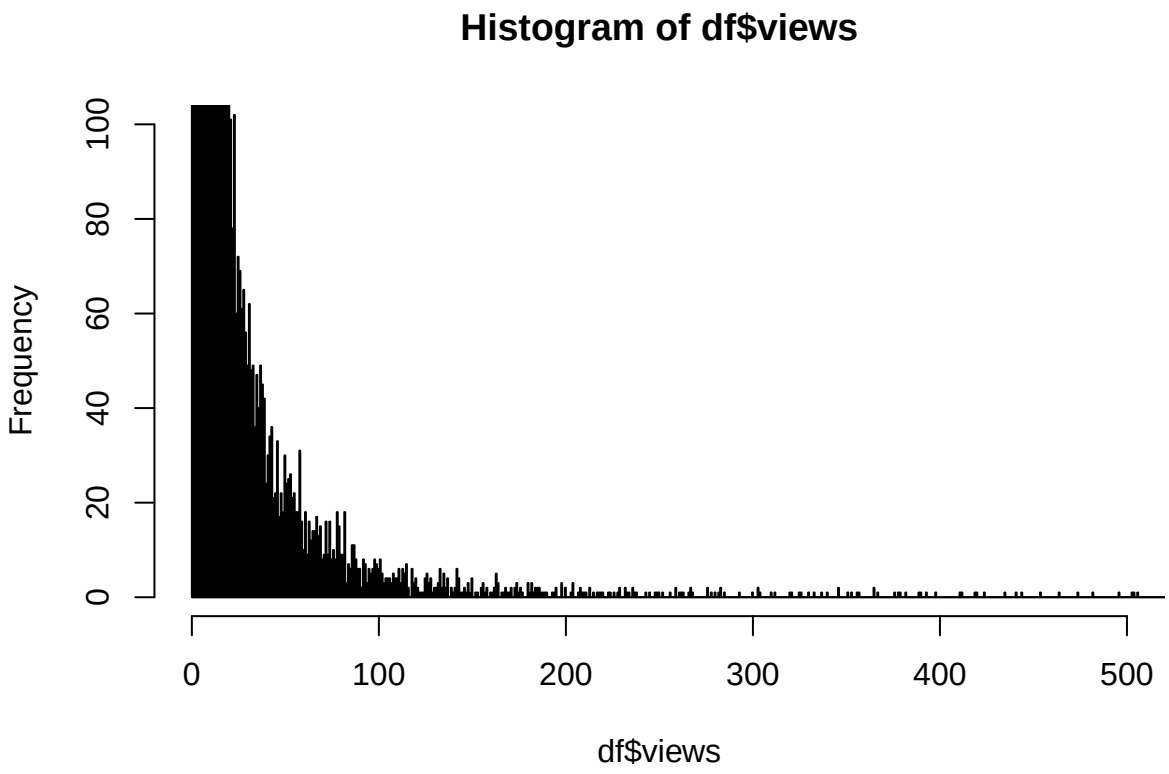

```
hist(df$views, breaks = 500, xlim = c(0, 5000))
```



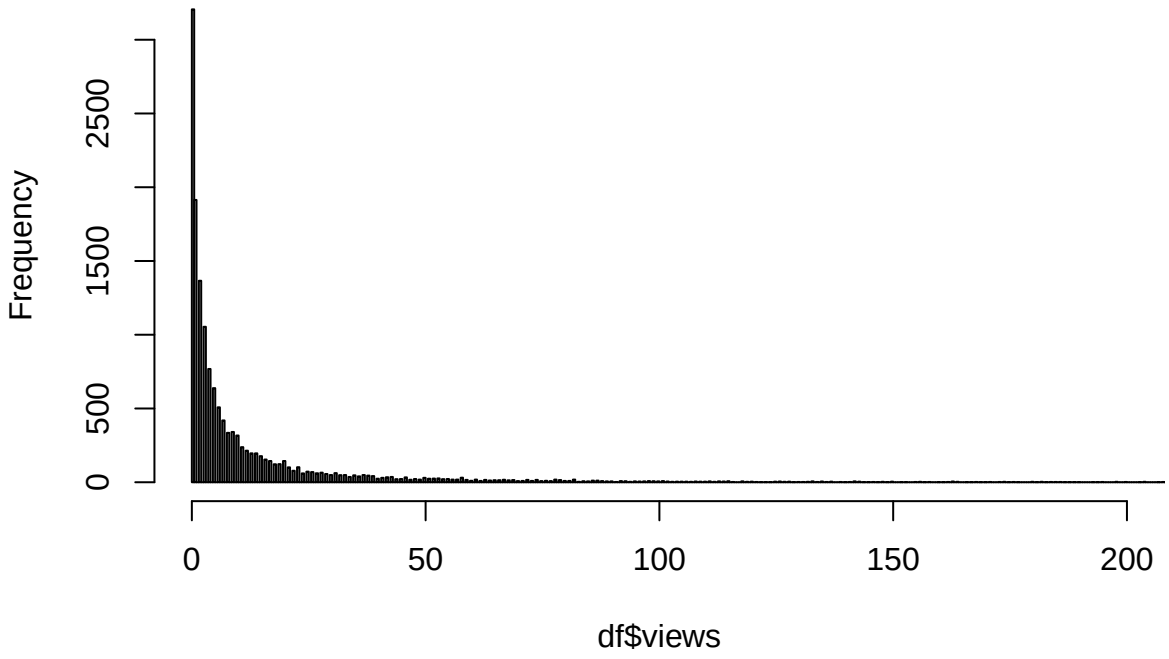

```r
hist(df$views, breaks = 5000, xlim = c(0, 500))
```




```
hist(df$views, breaks = 5000, xlim = c(0, 500), ylim = c(0, 100))
```




```r
hist(df$views, breaks = 5000, xlim = c(0, 200))
```




```
max(df$views) #max = 13079
```

```
## [1] 3234
```

```
df[which.max(df$views),]
```

```
##      period accessionyear dateend                title
## 7159                1951    1883 Singer with a Glove
##                                                                 url
## 7159 https://www.harvardartmuseums.org/collections/object/228652
##                       division      century objectnumber views datebegin
## 7159 European and American Art 19th century      1951.68  3234      1873
##      culture imagepermissionlevel  rank     id
## 7159  French                     0 81037 228652
##                                          department
## 7159 Department of Paintings, Sculpture & Decorative Arts
```

```
summary(df$views)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00   18.46   12.00 3234.00
```

```
quantile(df$views, c(0:20)/20)
```

```
##   0%   5%  10%  15%  20%  25%  30%  35%  40%  45%  50%  55%  60%  65%  70%  75%
##    0    0    0    0    0    1    1    2    2    3    3    4    6    7    9   12
##  80%  85%  90%  95% 100%
##   16   22   35   64 3234
```

```r
sum(df$views > 100) #421
```

```
## [1] 400
```

EDA of culture Some weird stuff, like with "European?" multiple categories of Italian, Roman, Spanish, British; there is Flemish and Franco-Flemish and French Unidentified culture and unknown are also two categories "Graeco-Bactrian"
"Graeco-Roman" + "Greek" "Hellenistic"
"Hellenistic or Early Roman"

```r
# broader stroke, 21 categories
head(df$culture)
```

```
## [1] "German"   "Japanese" "American" "American" "British"  "American"
```

```r
allcnames <- unique(df$culture)
length(allcnames)
```

```
## [1] 162
```

```r
orderedcnames <- allcnames[order(allcnames)]
#orderedcnames
# first remove the question marks, kinda pointless here
df$culture <- gsub("[?]", "", df$culture)

# Make consistent different locations
df$culture <- gsub("^British.+", "British", df$culture)
df$culture <- gsub("^Italian.+", "Italian", df$culture)
df$culture <- gsub("^Roman.+", "Roman", df$culture)
df$culture <- gsub("^Spanish.+", "Spanish", df$culture)

df$culture[df$culture == "" | df$culture == "Unidentified culture"  | df$culture == "unknown"] <- "Unkn

length(unique(df$culture))
```

```
## [1] 133
```

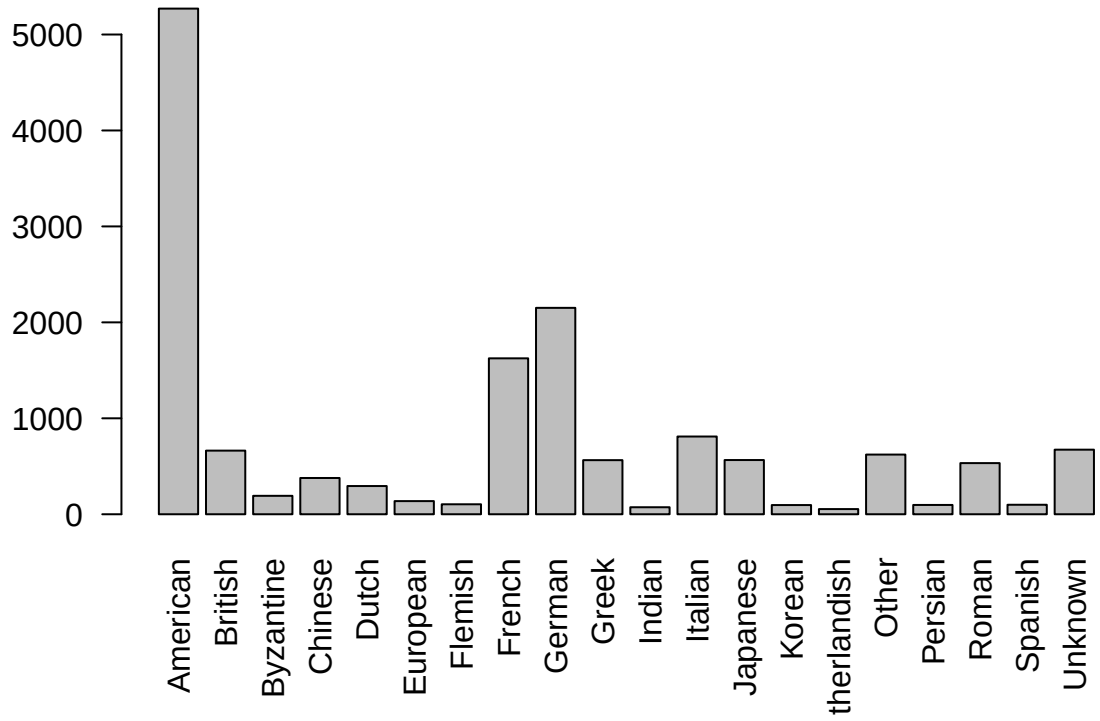
```r
# boarder stroke, 20 categories
df$fculture <- df$culture
culturerank <- names(which(table(df$culture) > 50))
length(culturerank)
```

```
## [1] 19
```

```r
df$fculture[!df$culture %in% culturerank] <- "Other"
table(df$fculture)
```

```
##
##      American       British     Byzantine       Chinese         Dutch
##          5270           663           192           378           294
##      European       Flemish        French        German         Greek
##           137           104          1625          2151           564
##        Indian       Italian      Japanese        Korean Netherlandish
##            73           810           565            96            54
##         Other       Persian         Roman       Spanish       Unknown
##           622            97           533            99           673
```

```r
df$fculture <- as.factor(df$fculture)
plot(df$fculture, las = 2)
```



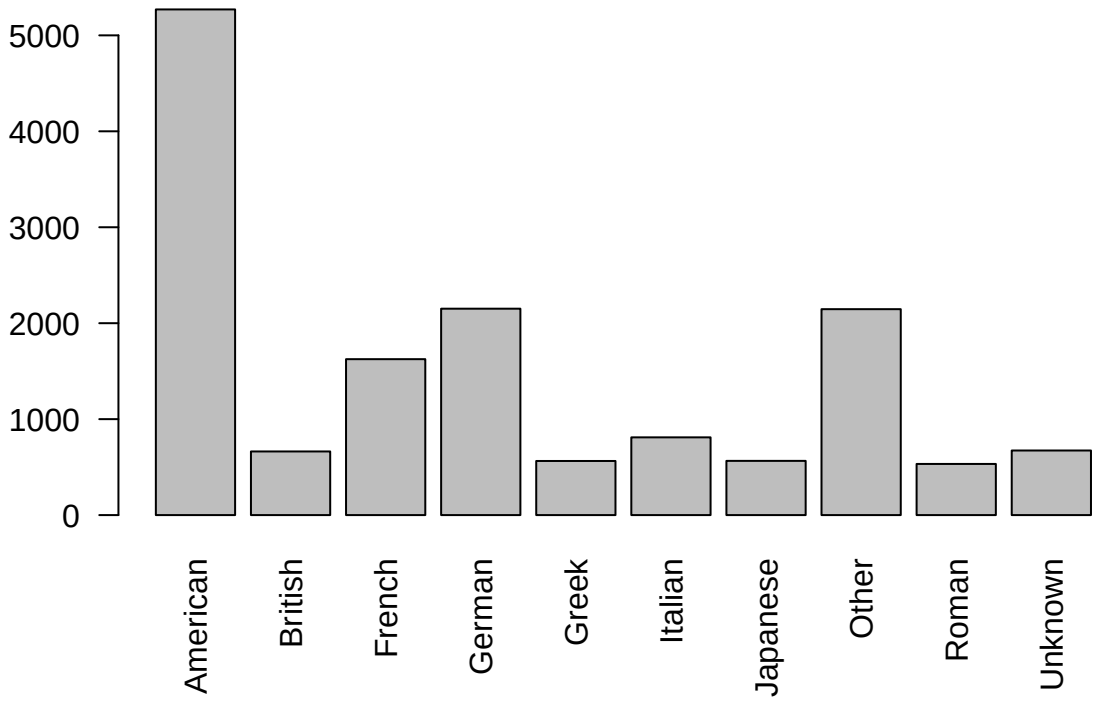

```
# smaller stroke, 10 categories
df$ffculture <- df$culture
culturerank <- names(which(table(df$culture) > 500))
length(culturerank)
```

```
## [1] 9
```

```
df$ffculture[!df$culture %in% culturerank] <- "Other"
table(df$ffculture)
```

```
##
## American  British   French   German    Greek  Italian Japanese    Other
##     5270      663     1625     2151      564      810      565     2146
##    Roman  Unknown
##      533      673
```

```
df$ffculture <- as.factor(df$ffculture)
plot(df$ffculture, las = 2)
```



EDA: department, division, century, period

```
head(df$department)
```

```
## [1] "Busch-Reisinger Museum"    "Department of Asian Art"
## [3] "Department of Photographs" "Department of Photographs"
## [5] "Department of Prints"      "Department of Photographs"
```

```
unique(df$department)
```

```
##  [1] "Busch-Reisinger Museum"
##  [2] "Department of Asian Art"
##  [3] "Department of Photographs"
##  [4] "Department of Prints"
##  [5] "Department of Drawings"
##  [6] "Department of Ancient and Byzantine Art & Numismatics"
##  [7] "Department of Paintings, Sculpture & Decorative Arts"
##  [8] "Straus Center for Conservation and Technical Studies"
##  [9] "Department of Islamic & Later Indian Art"
## [10] "Harvard University Portrait Collection"
## [11] "Department of American Paintings, Sculpture & Decorative Arts"
## [12] "Department of Modern & Contemporary Art"
## [13] "Archives"
## [14] "Center for the Technical Study of Modern Art"
## [15] "Harvard University Clock Collection"
```

```
df$department <- as.factor(df$department)
head(df$division)
```

```
## [1] "Modern and Contemporary Art" "Asian and Mediterranean Art"
## [3] "Modern and Contemporary Art" "Modern and Contemporary Art"
## [5] "European and American Art"   "Modern and Contemporary Art"
```

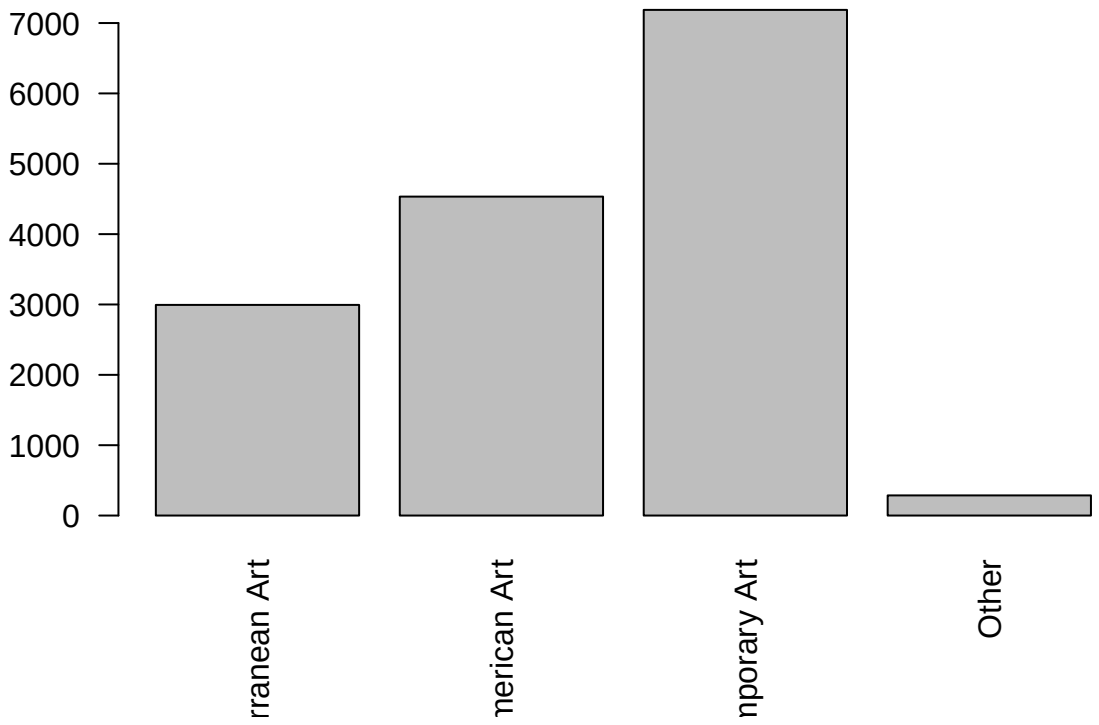
```
length(unique(df$division))
```

```
## [1] 4
```

```
unique(df$division)
```

```
## [1] "Modern and Contemporary Art" "Asian and Mediterranean Art"
## [3] "European and American Art"   ""
```

```
df$division[df$division == ""] <- "Other"
df$division <- as.factor(df$division)
plot(df$division, las = 2)
```




```
head(df$period)
```

```
## [1] ""                     "Edo period, 1615-1868" ""
## [4] ""                     ""                      ""
```

```
length(unique(df$period))
```

```
## [1] 172
```

```
head(df$century)
```

```
## [1] "20th century"      "18th-19th century" "20th century"
## [4] "20th century"      "19th century"      "20th century"
```

```r
length(unique(df$century))
```

```
## [1] 130
```

```r
unique(df$century)
```

```
##   [1] "20th century"
##   [2] "18th-19th century"
##   [3] "19th century"
##   [4] "19th-20th century"
##   [5] "18th century"
##   [6] "7th-6th century BCE"
##   [7] "21st century"
##   [8] "14th century"
##   [9] ""
##  [10] "17th century"
##  [11] "16th century"
##  [12] "14th-15th century"
##  [13] "12th century"
##  [14] "15th century"
##  [15] "5th century BCE"
##  [16] "2nd century BCE"
##  [17] "2nd century CE"
##  [18] "4th century CE"
##  [19] "11th-12th century"
##  [20] "15th century BCE"
##  [21] "6th century"
##  [22] "5th century"
##  [23] "17th-18th century"
##  [24] "3rd century BCE"
##  [25] "10th century"
##  [26] "4th century BCE"
##  [27] "16th-17th century"
##  [28] "5th-3rd century BCE"
##  [29] "2nd-3rd century CE"
##  [30] "16th-11th century BCE"
##  [31] "1st century BCE"
##  [32] "1st millennium BCE-1st millenium CE"
##  [33] "3rd century CE"
##  [34] "17th-19th century"
##  [35] "1st century BCE-1st century CE"
##  [36] "7th century"
##  [37] "5th-6th century"
##  [38] "1st century CE"
##  [39] "7th-8th century"
##  [40] "9th-10th century"
##  [41] "6th century BCE"
##  [42] "3rd millennium BCE"
##  [43] "4th-3rd century BCE"
##  [44] "2nd-1st century BCE"
##  [45] "8th century"
##  [46] "3rd-7th century"
##  [47] "15th-16th century"
##  [48] "16th-14th century BCE"
```

```
##  [49] "6th-5th century BCE"
##  [50] "12th-13th century"
##  [51] "5th-3rd millennium BCE"
##  [52] "9th-11th century"
##  [53] "4th-2nd millennium BCE"
##  [54] "2nd millennium BCE"
##  [55] "1st-4th century CE"
##  [56] "5th-4th century BCE"
##  [57] "13th century"
##  [58] "7th century BCE"
##  [59] "9th-8th century BCE"
##  [60] "10th-7th century BCE"
##  [61] "1st-3rd century CE"
##  [62] "13th-14th century"
##  [63] "9th century"
##  [64] "11th century"
##  [65] "1st-2nd century CE"
##  [66] "10th-13th century"
##  [67] "1st century BCE-3rd century CE"
##  [68] "17th-20th century"
##  [69] "3rd-2nd century BCE"
##  [70] "20th-21st century"
##  [71] "11th-13th century"
##  [72] "10th-8th century BCE"
##  [73] "4th-3rd millennium BCE"
##  [74] "5th-4th millennium BCE"
##  [75] "3rd-1st century BCE"
##  [76] "10th-9th century BCE"
##  [77] "8th-7th century BCE"
##  [78] "8th-9th century"
##  [79] "Unidentified century"
##  [80] "6th-7th century"
##  [81] "5th-7th century"
##  [82] "1st-2nd millennium CE"
##  [83] "11th-10th century BCE"
##  [84] "14th-16th century"
##  [85] "11th-15th century"
##  [86] "4th-5th century CE"
##  [87] "3rd century BCE-3rd century CE"
##  [88] "16th and 19th century"
##  [89] "7th-1st century BCE"
##  [90] "14th-12th century BCE"
##  [91] "6th millennium BCE"
##  [92] "3rd-2nd millennium BCE"
##  [93] "8th century BCE"
##  [94] "8th-5th century BCE"
##  [95] "15th-17th century"
##  [96] "12th-11th century BCE"
##  [97] "14th-11th century BCE"
##  [98] "4th millennium BCE"
##  [99] "12th-14th century"
## [100] "16th-13th century BCE"
## [101] "1st millennium CE"
## [102] "10th-12th century"
```

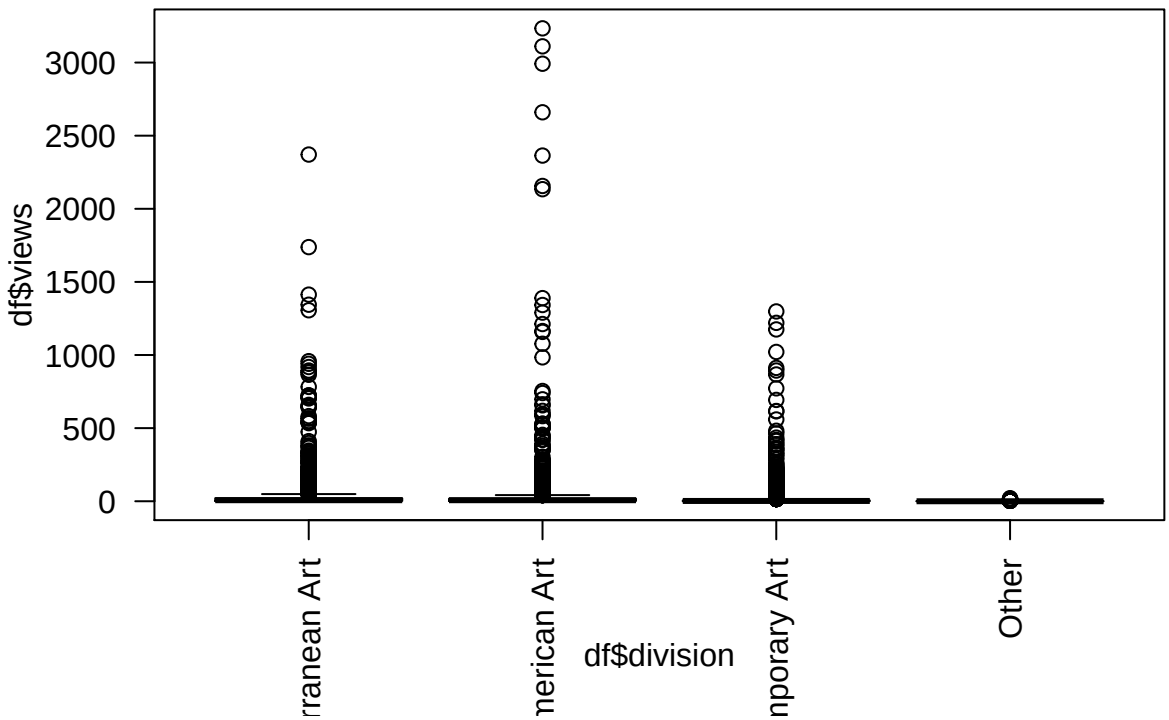
```
## [103] "10th century BCE"
## [104] "15th-13th century BCE"
## [105] "14th-17th century"
## [106] "18th and 19th centuries"
## [107] "14th-8th century BCE"
## [108] "2nd-4th century CE"
## [109] "4th-1st century BCE"
## [110] "10th-11th century"
## [111] "10th-14th century"
## [112] "8th-10th century"
## [113] "9th-7th century BCE"
## [114] "1st-5th century CE"
## [115] "11th-8th century BCE"
## [116] "3rd century BCE-1st century CE"
## [117] "4th-2nd century BCE"
## [118] "6th-4th century BCE"
## [119] "16th-18th century"
## [120] "13th-11th century BCE"
## [121] "1st century BCE-2nd century CE"
## [122] "8th-6th century BCE"
## [123] "5th millennium BCE"
## [124] "6th-5th millennium BCE"
## [125] "14th-13th century BCE"
## [126] "3rd-4th century CE"
## [127] "18th-20th century"
## [128] "1st millennium BCE"
## [129] "8th-2nd century BCE"
## [130] "9th century BCE"
```

```
head(df$dateend)
```

```
## [1] 1930 1832 1968 1956    0 1955
```

EDA: Intersections

```
plot(df$views ~df$division, las = 2)
```



oh god, the dated stuff will be very annoying to clean

```
library(stringr)
head(df$dated)

# get rid of those unknown stuff
uuu <- str_detect(df$dated, "^[Uu]")
allunknown <- c(unique(df$dated[uuu]), "")
sum(df$dated == "Unknown")
df$dated[df$dated %in% allunknown] <- "Unknown"

# lowkey lumping all the BCE together, as not too many of them
bce <- str_detect(df$dated, "BCE")
sum(df$dated[bce])
df$dated[bce] <- "BCE"

# then hopefully can lump by century GG
# first work with the best case scenario LOL
# thank u random guy: https://gist.github.com/micstr/69a64fbd0f5635094a53
IsDate <- function(mydate, date.format = "%m/%d/%y") {
  tryCatch(!is.na(as.Date(mydate, date.format)),
           error = function(err) {FALSE})
}

nicedates <- IsDate(df$dated)
sum(nicedates)
a = c("10/2/2020", 13, "asd")
```

```r
a[1] <- as.Date(a[1])
substr(a, 1, 2)
head(df$dated)
unique(df$dated)
sum(df$dated == "Unknown")
```