la2

Hariprasad H N

2024-01-30

install.packages(c("tidyverse", "dplyr", "tidyr", "readr", "knitr", "jsonlite", "lubridate", "stringr", "rmarkdown")) install.packages("ggplot2")

install.packages('tinytex') tinytex::install_tinytex()

```
setwd("C:/Users/harsh/Music/edal2/archive")

us_videos <- read.csv("USvideos.csv")
gb_videos <- read.csv("GBvideos.csv")
gb_comments <- read.csv("GBcomments.csv")
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : EOF within quoted string
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : embedded nul(s) found in input
```

```
us_categories <- jsonlite::fromJSON("US_category_id.json")
gb_categories <- jsonlite::fromJSON("GB_category_id.json")

#data inspection
str(us_videos)
```

```
## 'data.frame':    8004 obs. of  11 variables:
##  $ video_id      : chr  "XpVt6Z1Gjjo" "K4wEI5zhHB0" "cLdxuaxaQwc" "WYYvHb03Eog" ...
##  $ title         : chr  "1 YEAR OF VLOGGING -- HOW LOGAN PAUL CHANGED YOUTUBE FOREVER!" "iPhone X - I
##  $ channel_title : chr  "Logan Paul Vlogs" "Apple" "PewDiePie" "The Verge" ...
##  $ category_id   : chr  "24" "28" "22" "28" ...
##  $ tags          : chr  "logan paul vlog|logan paul|logan|paul|olympics|logan paul youtube|vlog|daily
##  $ views         : int  4394029 7860119 5845909 2642103 1168130 1311445 666169 1728614 1338533 10568
##  $ likes         : int  320053 185853 576597 24975 96666 34507 9985 74062 69687 29943 ...
##  $ dislikes      : int  5931 26679 39774 4542 568 544 297 2180 678 878 ...
##  $ comment_total : chr  "46245" "0" "170708" "12829" ...
##  $ thumbnail_link: chr  "https://i.ytimg.com/vi/XpVt6Z1Gjjo/default.jpg" "https://i.ytimg.com/vi/K4wI
##  $ date          : chr  "13.09" "13.09" "13.09" "13.09" ...
```

```
head(us_videos)
```

```
##      video_id                                                        title
## 1 XpVt6Z1Gjjo 1 YEAR OF VLOGGING -- HOW LOGAN PAUL CHANGED YOUTUBE FOREVER!
```

```
## 2 K4wEI5zhHB0                       iPhone X - Introducing iPhone X - Apple
## 3 cLdxuaxaQwc                                                   My Response
## 4 WYYvHb03Eog                                  Apple iPhone X first look
## 5 sjlHnJvXdQs                                          iPhone X (parody)
## 6 cMKX2tE5Luk              The Disaster Artist | Official Trailer HD | A24
##      channel_title category_id
## 1 Logan Paul Vlogs          24
## 2            Apple          28
## 3        PewDiePie          22
## 4        The Verge          28
## 5       jacksfilms          23
## 6              A24           1
##
## 1                                                                                  logan paul
## 2
## 3
## 4
## 5
## 6 a24|a24 films|a24 trailers|independent films|trailer|HD|official|movie|film|a24 movies|oscar winne
##     views  likes dislikes comment_total
## 1 4394029 320053     5931         46245
## 2 7860119 185853    26679             0
## 3 5845909 576597    39774        170708
## 4 2642103  24975     4542         12829
## 5 1168130  96666      568          6666
## 6 1311445  34507      544          3040
##                                 thumbnail_link  date
## 1 https://i.ytimg.com/vi/XpVt6Z1Gjjo/default.jpg 13.09
## 2 https://i.ytimg.com/vi/K4wEI5zhHB0/default.jpg 13.09
## 3 https://i.ytimg.com/vi/cLdxuaxaQwc/default.jpg 13.09
## 4 https://i.ytimg.com/vi/WYYvHb03Eog/default.jpg 13.09
## 5 https://i.ytimg.com/vi/sjlHnJvXdQs/default.jpg 13.09
## 6 https://i.ytimg.com/vi/cMKX2tE5Luk/default.jpg 13.09
```

```
str(us_categories)
```

```
## List of 3
##  $ kind : chr "youtube#videoCategoryListResponse"
##  $ etag : chr "\"m2yskBQFythfE4irbTIeOgYYfBU/S730Ilt-Fi-emsQJvJAAShlR6hM\""
##  $ items:'data.frame':   32 obs. of  4 variables:
##   ..$ kind   : chr [1:32] "youtube#videoCategory" "youtube#videoCategory" "youtube#videoCategory" "y
##   ..$ etag   : chr [1:32] "\"m2yskBQFythfE4irbTIeOgYYfBU/Xy1mB4_yLrHy_BmKmPBggty2mZQ\"" "\"m2yskBQFy
##   ..$ id     : chr [1:32] "1" "2" "10" "15" ...
##   ..$ snippet:'data.frame':  32 obs. of  3 variables:
##   .. ..$ channelId : chr [1:32] "UCBR8-60-B28hp2BmDPdntcQ" "UCBR8-60-B28hp2BmDPdntcQ" "UCBR8-60-B28h
##   .. ..$ title     : chr [1:32] "Film & Animation" "Autos & Vehicles" "Music" "Pets & Animals" ...
##   .. ..$ assignable: logi [1:32] TRUE TRUE TRUE TRUE TRUE FALSE ...
```

```
#for data summery
summary(us_videos)
```

```
##    video_id            title           channel_title      category_id
##  Length:8004        Length:8004        Length:8004        Length:8004
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      tags               views              likes            dislikes
##  Length:8004        Min.   :       0   Min.   :      0   Min.   :     0
##  Class :character   1st Qu.:   96891   1st Qu.:   1900   1st Qu.:    68
##  Mode  :character   Median :  308316   Median :   8645   Median :   273
##                     Mean   :  938893   Mean   :  34501   Mean   :  1786
##                     3rd Qu.:  958876   3rd Qu.:  30214   3rd Qu.:  1014
##                     Max.   :41500672   Max.   :2010366   Max.   :318404
##                     NA's   :4          NA's   :4         NA's   :4
##  comment_total      thumbnail_link         date
##  Length:8004        Length:8004        Length:8004
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
us_category_info <- us_categories$items %>%
                    select(category_id = id)
us_videos <- merge(us_videos, us_category_info, by="category_id")

str(us_categories$items[1:3])
```

```
## 'data.frame':    32 obs. of  3 variables:
##  $ kind: chr  "youtube#videoCategory" "youtube#videoCategory" "youtube#videoCategory" "youtube#videoC
##  $ etag: chr  "\"m2yskBQFythfE4irbTIeOgYYfBU/Xy1mB4_yLrHy_BmKmPBggty2mZQ\"" "\"m2yskBQFythfE4irbTIeOg
##  $ id  : chr  "1" "2" "10" "15" ...
```

```
#data clean
#check missing values
sum(is.na(us_videos))
```

```
## [1] 0
```

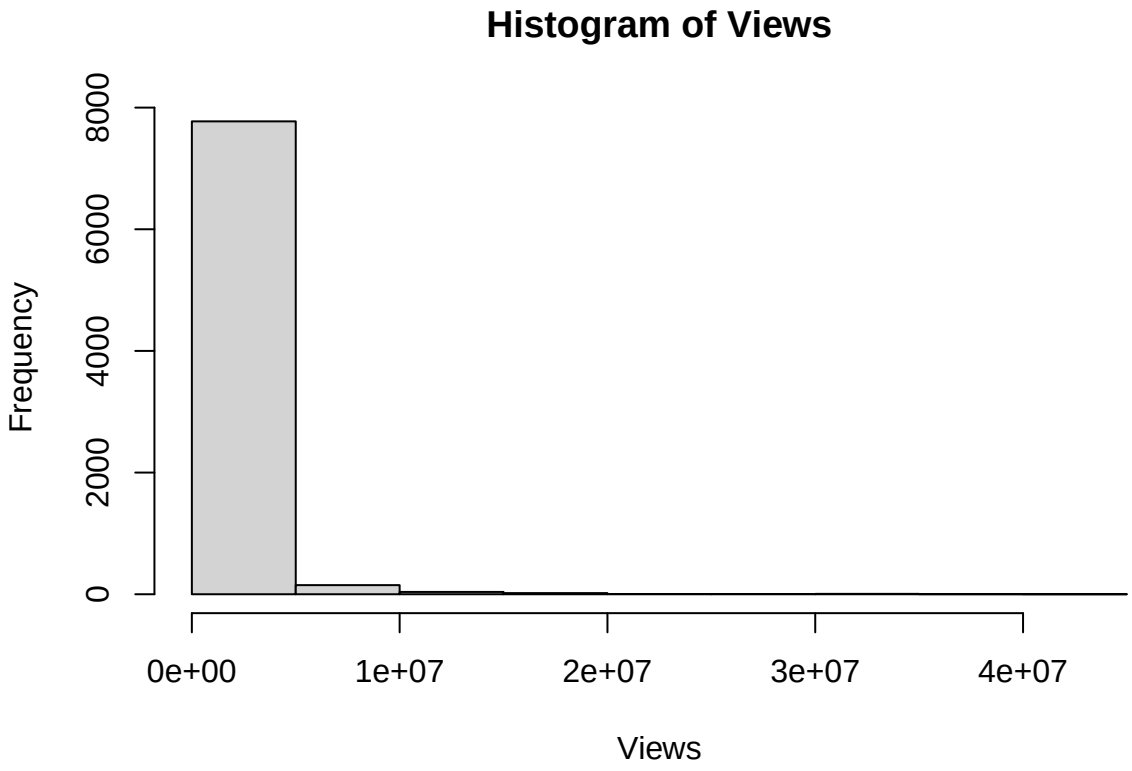
```r
#Check the structure and types of each column:
str(us_videos)
```

```
## 'data.frame':    7998 obs. of  11 variables:
##  $ category_id   : chr  "1" "1" "1" "1" ...
##  $ video_id      : chr  "fUjicxMPDzs" "rHfyvSgvgoo" "zgLtEob6X-Q" "xNddRhpx5tA" ...
##  $ title         : chr  "Pacific Rim Uprising - Official Trailer (HD)" "Game Revealed: Season 7 Episo
##  $ channel_title : chr  "Legendary" "GameofThrones" "Screen Junkies" "CinemaSins" ...
##  $ tags          : chr  "[none]" "Game of Thrones|Iron Throne|GOT|White Walker|Khalessi|Winter is Co
##  $ views         : int  5741857 216713 1056891 969749 696361 1764868 971242 903015 571327 13373 ...
##  $ likes         : int  49038 5329 29943 23072 16946 46524 26435 22173 8966 680 ...
##  $ dislikes      : int  6865 60 878 447 469 1183 748 422 822 20 ...
##  $ comment_total : chr  "14773" "471" "4046" "3871" ...
##  $ thumbnail_link: chr  "https://i.ytimg.com/vi/fUjicxMPDzs/default.jpg" "https://i.ytimg.com/vi/rHfy
##  $ date          : chr  "12.10" "15.09" "13.09" "18.10" ...
```

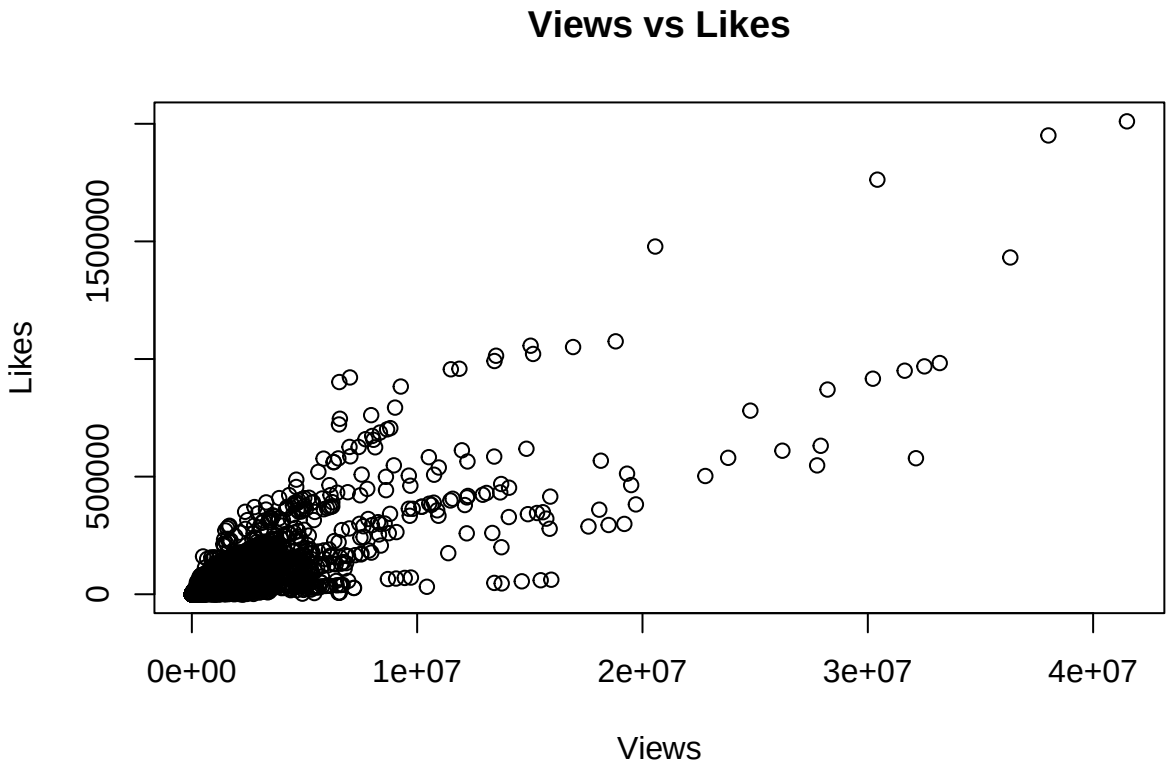
```r
#Remove rows with NA values (if that's your chosen approach):
us_videos <- na.omit(us_videos)

#Exploratory Data Analysis

#Summary Statistics:
summary(us_videos)
```

```
##  category_id          video_id           title           channel_title
##  Length:7998        Length:7998        Length:7998        Length:7998
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      tags               views              likes            dislikes
##  Length:7998        Min.   :       0   Min.   :      0   Min.   :     0
##  Class :character   1st Qu.:   96900   1st Qu.:   1902   1st Qu.:    68
##  Mode  :character   Median :  308612   Median :   8650   Median :   273
##                     Mean   :  939102   Mean   :  34509   Mean   :  1781
##                     3rd Qu.:  959513   3rd Qu.:  30221   3rd Qu.:  1013
##                     Max.   :41500672   Max.   :2010366   Max.   :318404
##  comment_total      thumbnail_link         date
##  Length:7998        Length:7998        Length:7998
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
hist(us_videos$views, main="Histogram of Views", xlab="Views")
```




```
plot(us_videos$views, us_videos$likes, main="Views vs Likes", xlab="Views", ylab="Likes")
```




```
# Data Visualizatio
# Load necessary libraries
library(tidyverse)
library(jsonlite)
```

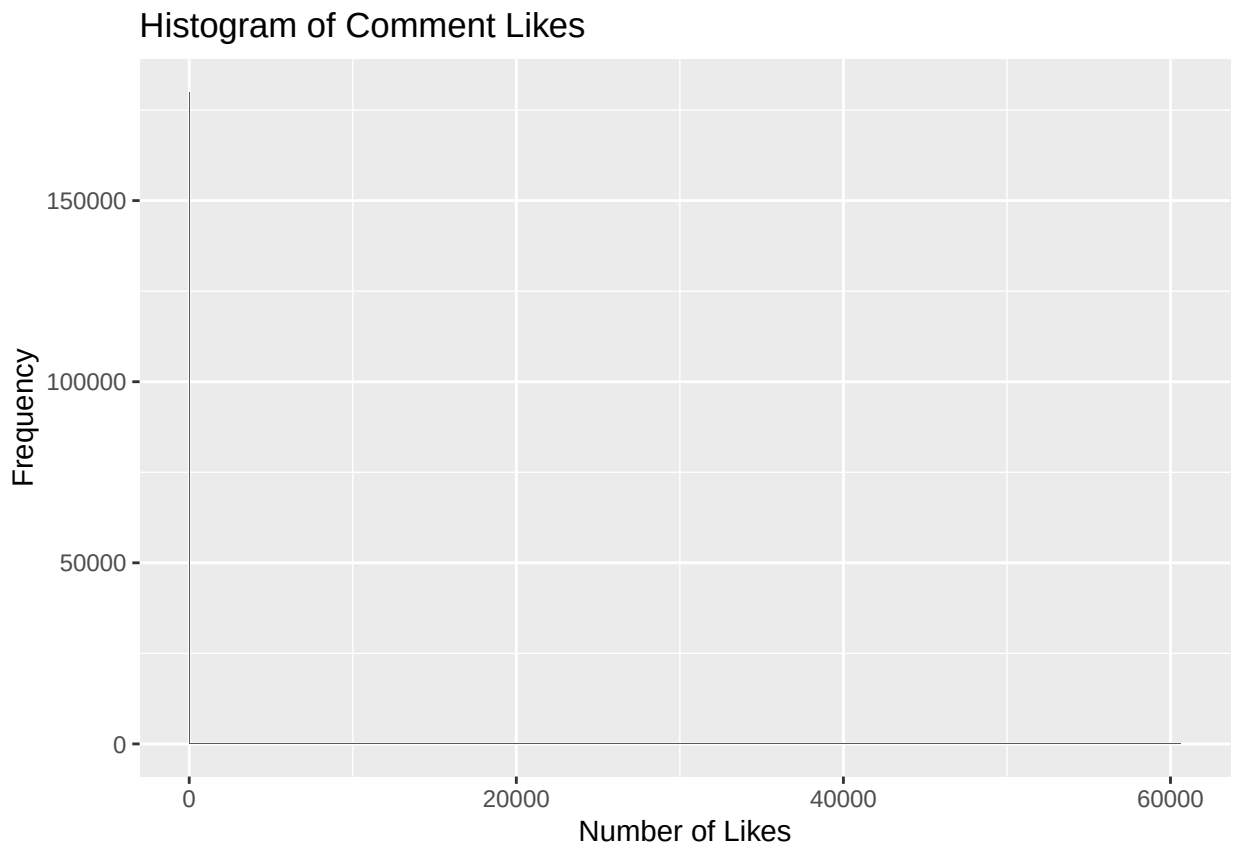
```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(ggplot2)

# Assuming gb_comments is already loaded and cleaned up
# Histogram of likes for the GB comments dataset
library(ggplot2)
gb_comments$likes <- as.numeric(gb_comments$likes)
```

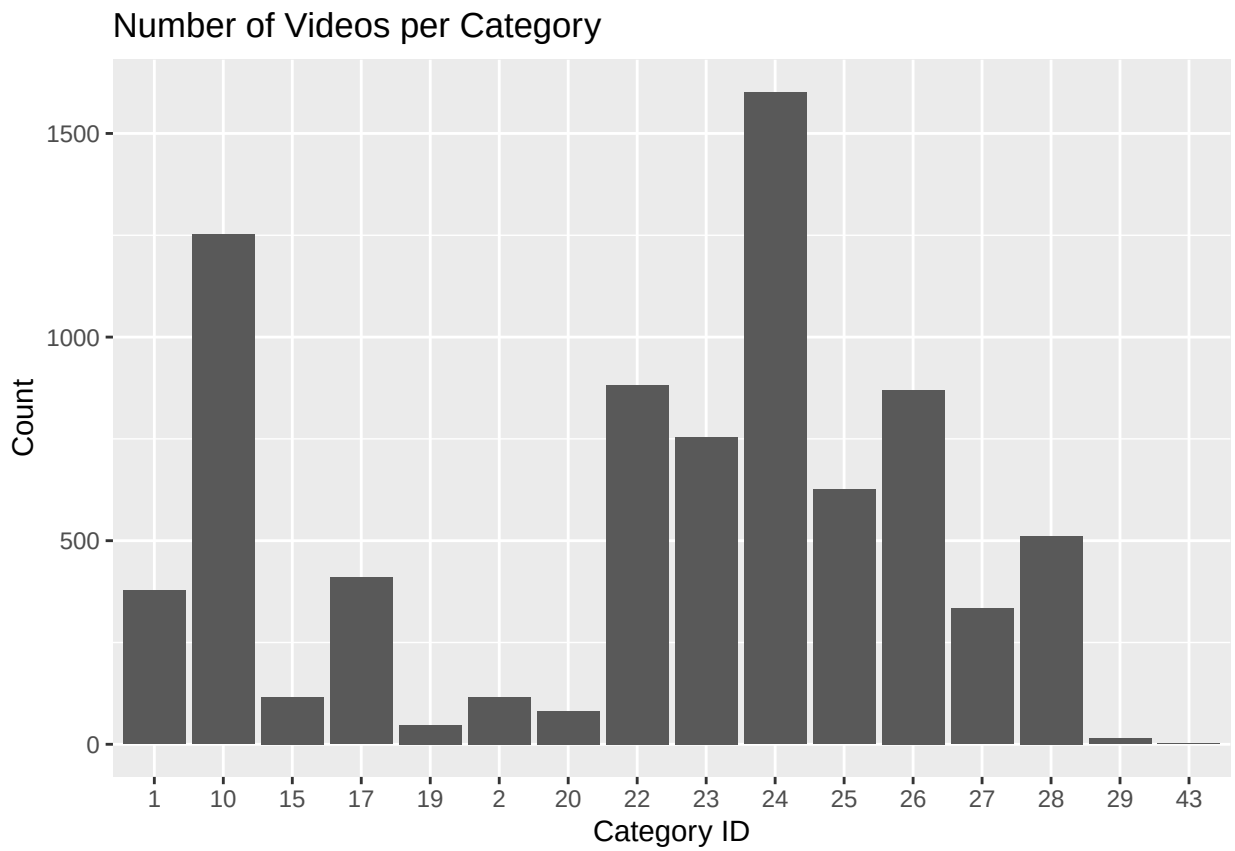
```
## Warning: NAs introduced by coercion
```

```
gb_comments$likes <- as.numeric(as.character(gb_comments$likes))
ggplot(gb_comments, aes(x=likes)) +
  geom_histogram(binwidth=1) +
  labs(title="Histogram of Comment Likes", x="Number of Likes", y="Frequency")
```

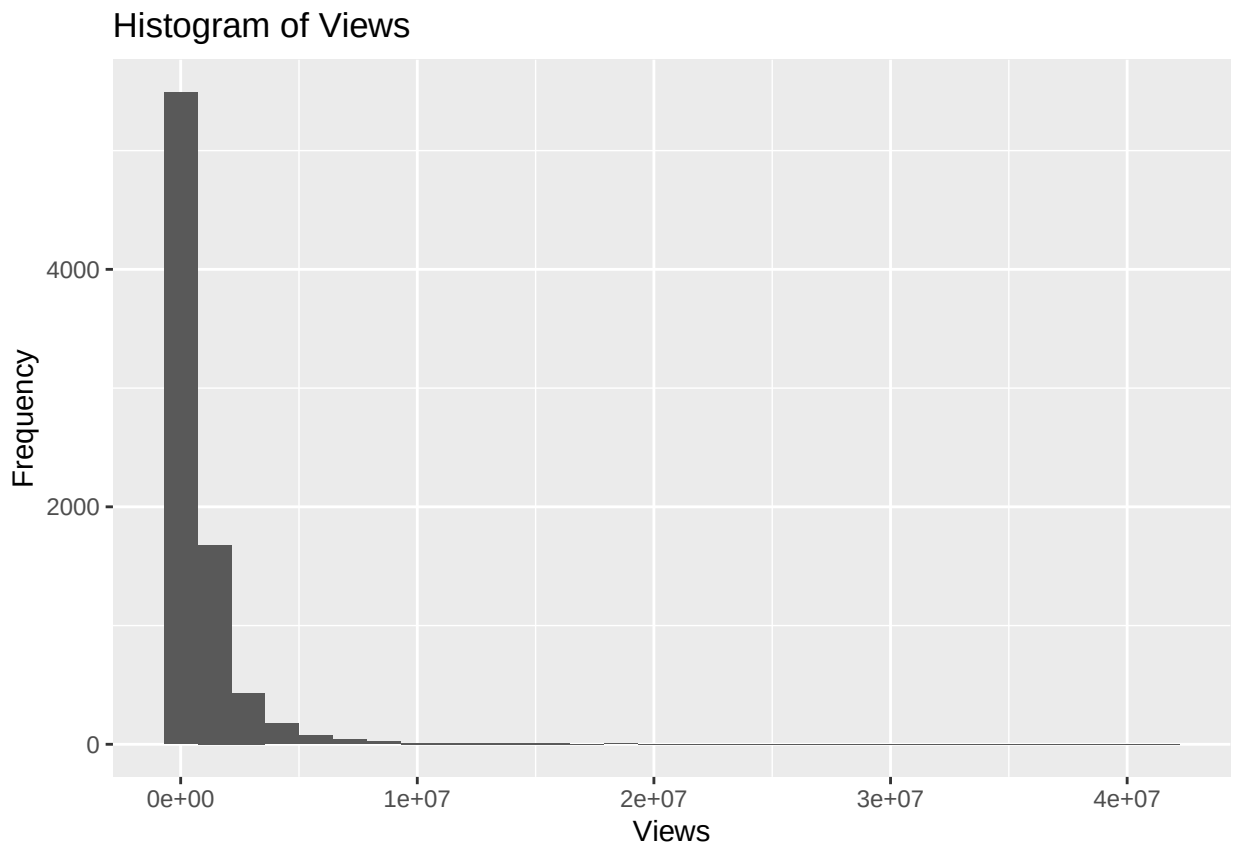
```
## Warning: Removed 47985 rows containing non-finite values (`stat_bin()`).
```



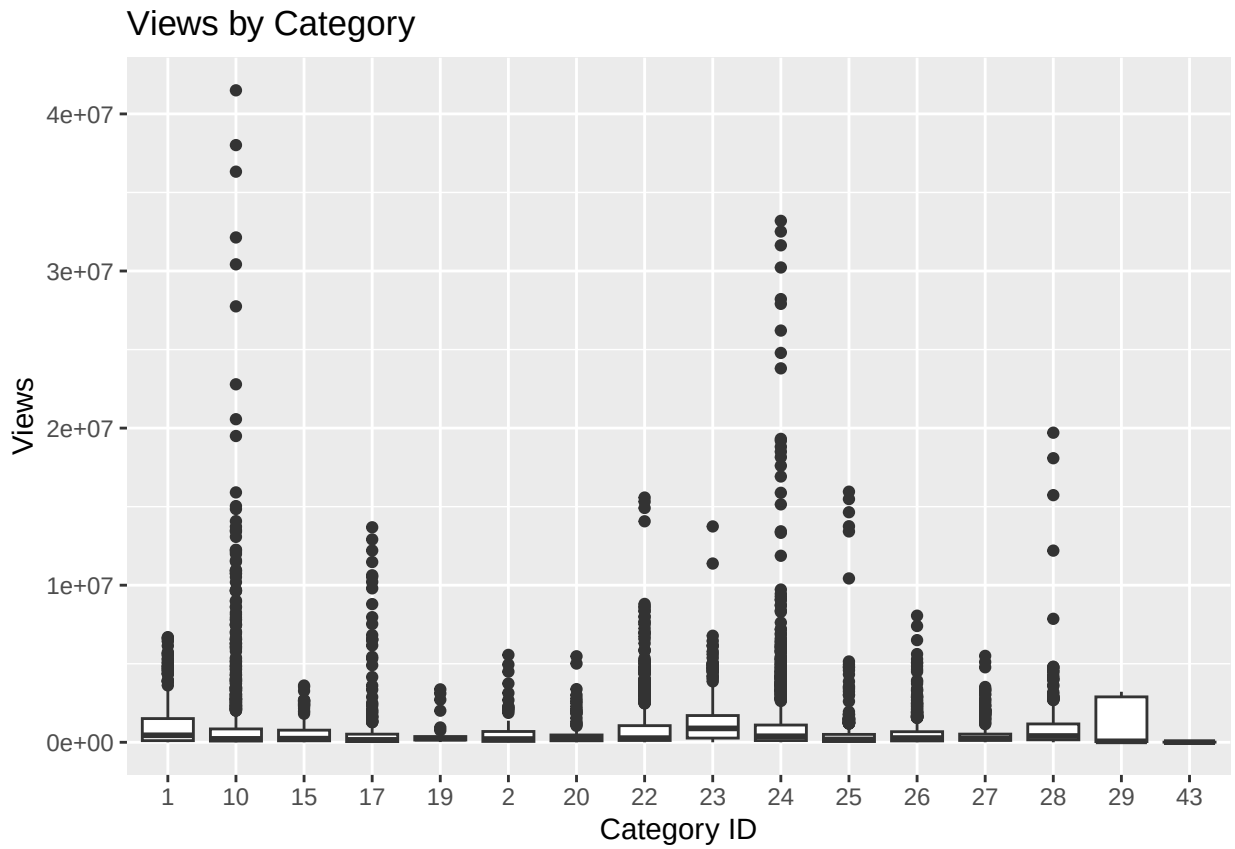

```
ggplot(us_videos, aes(x=factor(category_id))) +
  geom_bar() +
  labs(title="Number of Videos per Category", x="Category ID", y="Count")
```




```
ggplot(us_videos, aes(x=views)) +
  geom_histogram(bins=30) +
  labs(title="Histogram of Views", x="Views", y="Frequency")
```

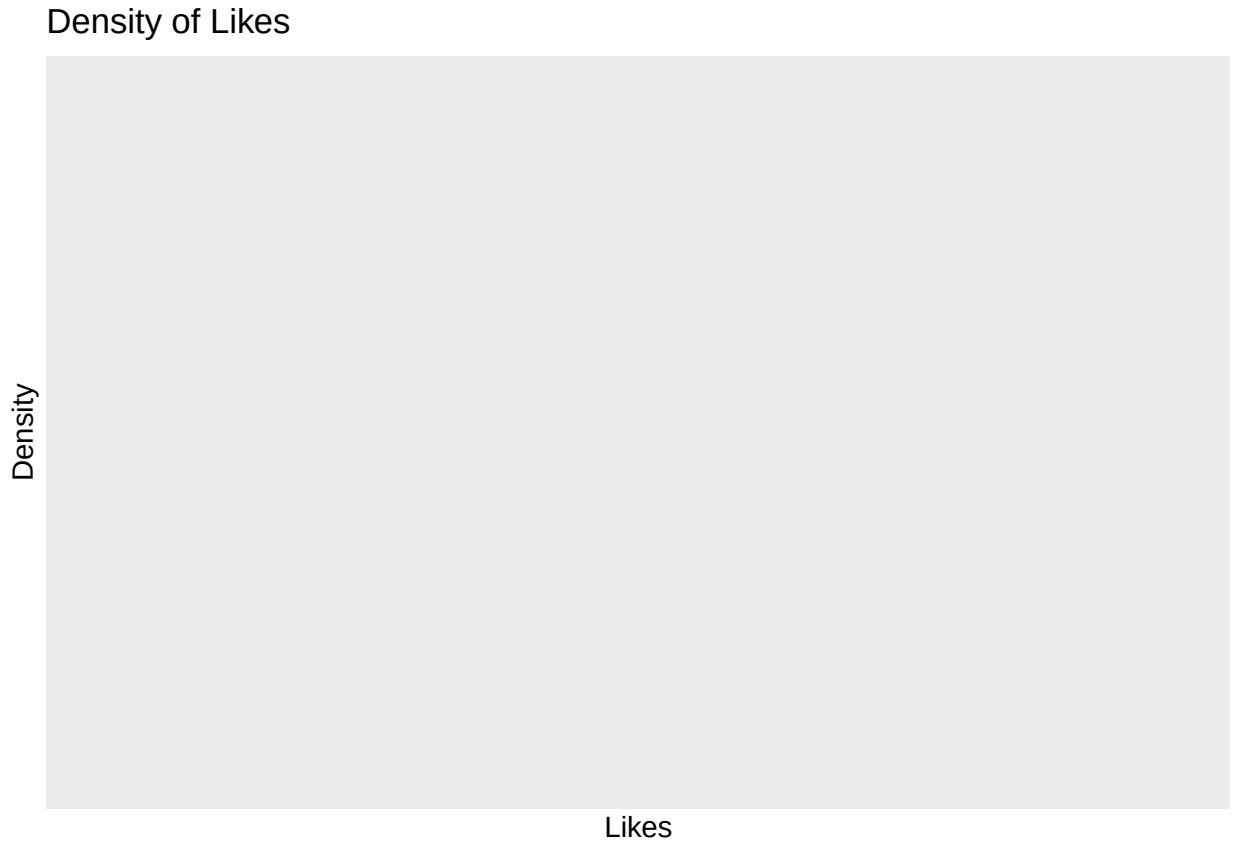



```
ggplot(us_videos, aes(x=factor(category_id), y=views)) +
  geom_boxplot() +
  labs(title="Views by Category", x="Category ID", y="Views")
```

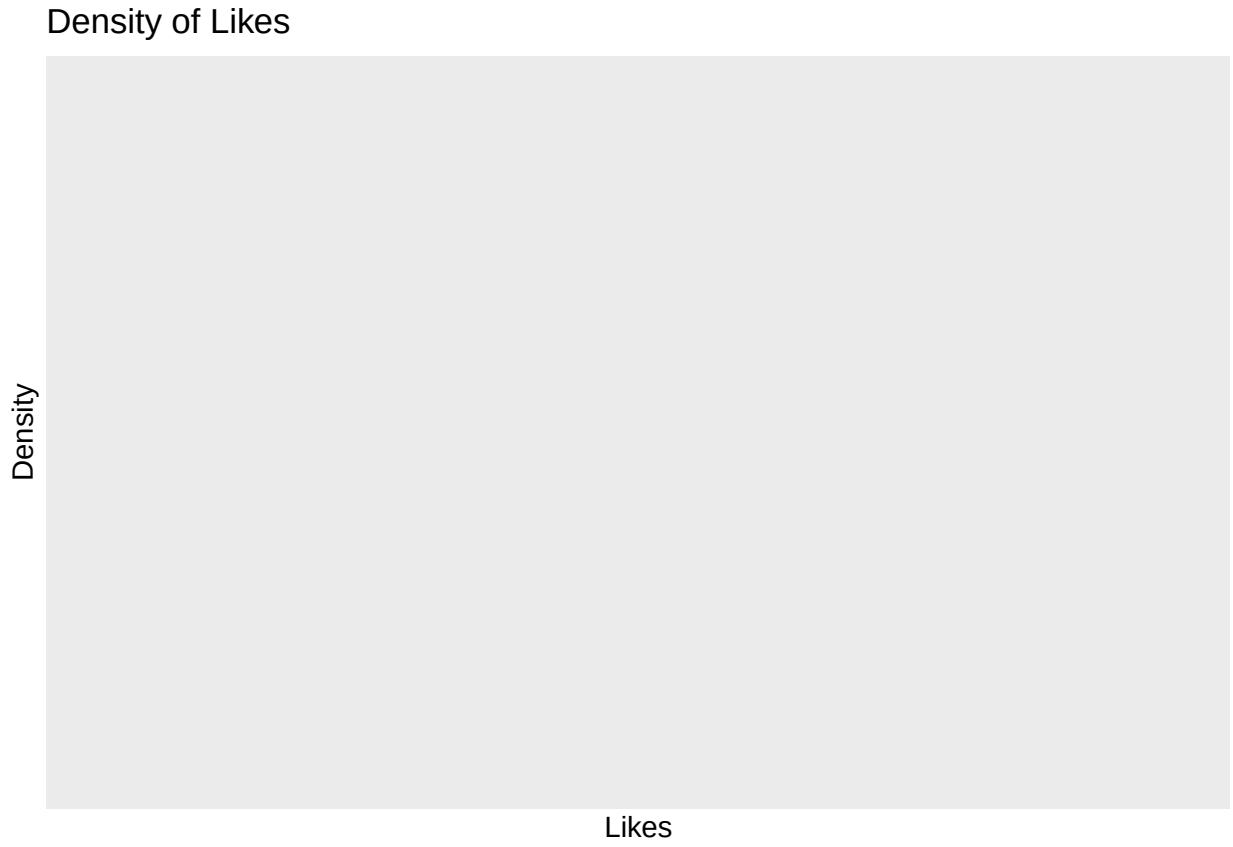



```r
# Convert the date column to a Date object before plotting
us_videos$date <- as.Date(us_videos$date, format = "%d-%m-%yt")  # Replace 'your_date_format' with your
us_videos <- na.omit(us_videos)  # Removes rows with NA
# or handle infinite values if any
us_videos <- us_videos[is.finite(us_videos$likes), ]


ggplot(us_videos, aes(x=date, y=likes)) +
  geom_line() +
  labs(title="Likes Over Time", x="Date", y="Likes")
```


```
ggplot(us_videos, aes(x=likes)) +
  geom_density() +
  labs(title="Density of Likes", x="Likes", y="Density")
```

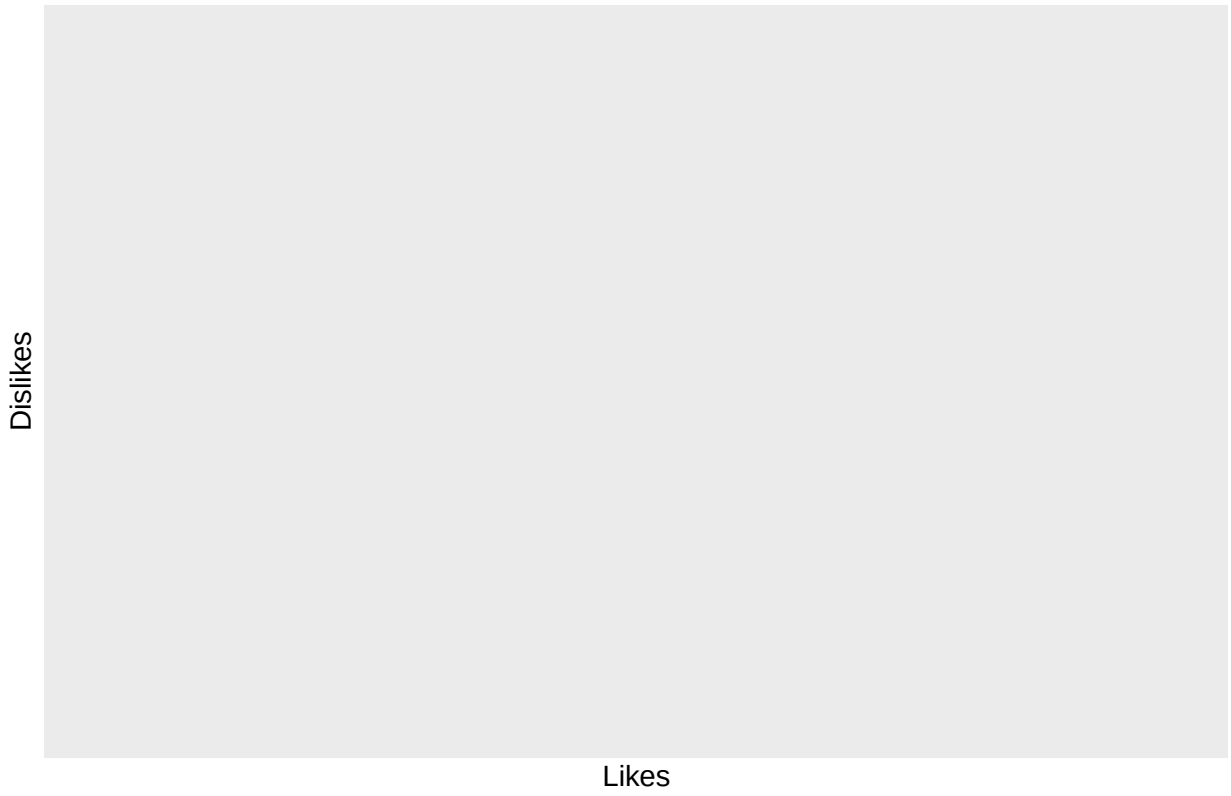



```
ggplot(us_videos, aes(x=likes)) +
  geom_density() +
  labs(title="Density of Likes", x="Likes", y="Density")
```




```
ggplot(us_videos, aes(x=factor(category_id), y=likes)) +
  geom_violin() +
  labs(title="Likes by Category", x="Category ID", y="Likes")
```


```
ggplot(us_videos, aes(x=likes, y=dislikes)) +
  geom_point() +
  labs(title="Likes vs. Dislikes", x="Likes", y="Dislikes")
```


Likes vs. Dislikes
Dislikes
Likes