Final project - DAO THI MINH NGOC

YouTube Trending Videos Dataset Analysis

Source: https://www.kaggle.com/datasets/thedevastator/youtube-trending-videos-dataset

1. Data preparation
   1.1. Data inspection

```
library(ggplot2)
df <- read.csv("/Users/macbook/Downloads/youtube.csv", header = TRUE)
str(df)
```

```
'data.frame':    161470 obs. of  16 variables:
 $ index              : int  0 1 2 3 4 5 6 7 8 9 ...
 $ trending_date      : chr  "17.14.11" "17.14.11" "17.14.11" "17.14.11" ...
 $ title              : chr  "WE WANT TO TALK ABOUT OUR MARRIAGE" "The Trump Presidency: Last Week Tonight with John Oliver (HBO)" "Racist Superman | Rudy Mancuso, King Bach & Lele Pons" "Nickelback Lyrics: Real or Fake?" ...
 $ channel_title      : chr  "CaseyNeistat" "LastWeekTonight" "Rudy Mancuso" "Good Mythical Morning" ...
 $ category_id        : int  22 24 23 24 24 28 24 28 1 25 ...
 $ publish_date       : chr  "13/11/2017" "13/11/2017" "12/11/2017" "13/11/2017" ...
 $ time_frame         : chr  "17:00 to 17:59" "7:00 to 7:59" "19:00 to 19:59" "11:00 to 11:59" ...
 $ published_day_of_week : chr  "Monday" "Monday" "Sunday" "Monday" ...
 $ publish_country    : chr  "US" "US" "US" "US" ...
 $ views              : int  748374 2418783 3191434 343168 2095731 119180 2103417 817732 826059 256426 ...
 $ likes              : int  57527 97185 146033 10172 132235 9763 15993 23663 3543 12654 ...
 $ dislikes           : int  2966 6146 5339 666 1989 511 2445 778 119 1363 ...
 $ comment_count      : int  15954 12703 8181 2146 17518 1434 1970 3432 340 2368 ...
 $ comments_disabled  : chr  "False" "False" "False" "False" ...
 $ ratings_disabled   : chr  "False" "False" "False" "False" ...
 $ video_error_or_removed: chr  "False" "False" "False" "False" ...
```

```
View(df)
```

	index	trending_date	title	channel_title	category_id	publish_date	time_frame
1	0	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	13/11/2017	17:00 to 17:59
2	1	17.14.11	The Trump Presidency: Last Week Tonight with John ...	LastWeekTonight	24	13/11/2017	7:00 to 7:59
3	2	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Lele Po...	Rudy Mancuso	23	12/11/2017	19:00 to 19:59
4	3	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	13/11/2017	11:00 to 11:59
5	4	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	12/11/2017	18:00 to 18:59
6	5	17.14.11	2 Weeks with iPhone X	iJustine	28	13/11/2017	19:00 to 19:59
7	6	17.14.11	Roy Moore & Jeff Sessions Cold Open - SNL	Saturday Night Live	24	12/11/2017	5:00 to 5:59
8	7	17.14.11	5 Ice Cream Gadgets put to the Test	CrazyRussianHacker	28	12/11/2017	21:00 to 21:59
9	8	17.14.11	The Greatest Showman | Official Trailer 2 [HD] | 20th ...	20th Century Fox	1	13/11/2017	14:00 to 14:59
10	9	17.14.11	Why the rise of the robots wonâ€™t mean the end of ...	Vox	25	13/11/2017	13:00 to 13:59
11	10	17.14.11	Dion Lewis' 103-Yd Kick Return TD vs. Denver! | Can'...	NFL	17	13/11/2017	2:00 to 2:59

published_day_of_week	publish_country	views	likes	dislikes	comment_count	comments_disabled	ratings_disabled	video_error_or_removed
Monday	US	748374	57527	2966	15954	False	False	False
Monday	US	2418783	97185	6146	12703	False	False	False
Sunday	US	3191434	146033	5339	8181	False	False	False
Monday	US	343168	10172	666	2146	False	False	False
Sunday	US	2095731	132235	1989	17518	False	False	False
Monday	US	119180	9763	511	1434	False	False	False
Sunday	US	2103417	15993	2445	1970	False	False	False
Sunday	US	817732	23663	778	3432	False	False	False
Monday	US	826059	3543	119	340	False	False	False
Monday	US	256426	12654	1363	2368	False	False	False
Monday	US	81377	655	25	177	False	False	False

```
anyNA(df)
```

```
[1] FALSE
```

In conclusion:

- There are 161.470 videos with 16 variables included.
- The category still in numbered format.
- There are no missing values in this dataset.

1.2. Data cleansing

```
#The imported "df" dataframe has multiple entries for videos trending on different days. To ensure uniqueness, Create a dataframe that includes data only when each title trended.
```

```
youtube <- df[match(unique(df$title), df$title),]
```

```
#Change the category which still in numbered forma into character to make it more understadable.
```

```
youtube <- youtube %>%
  mutate(category_id = case_when(
    category_id == 1 ~ "Film and Animation",
    category_id == 2 ~ "Autos and Vehicles",
    category_id == 10 ~ "Music", category_id == 15 ~ "Pets and Animals",
    category_id == 17 ~ "Sports", category_id == 18 ~ "Short Movies",
    category_id == 19 ~ "Travel and Events", category_id == 20 ~ "Gaming",
    category_id == 22 ~ "People and Blogs", category_id == 23 ~ "Comedy",
    category_id == 24 ~ "Entertainment",
    category_id == 25 ~ "News and Politics",
    category_id == 26 ~ "Howto and Style",
    category_id == 27 ~ "Education",
    category_id == 28 ~ "Science and Technology",
    category_id == 29 ~ "Nonprofit and Activism",
    category_id == 30  ~ "Movies",
    category_id == 43 ~ "Shows",
    category_id == 44  ~ "Trailers",
    TRUE ~ as.character(category_id)
  ))
```

1.3. Adding new variables

#1 Period of the day

```
youtube <- youtube %>%
  mutate(periodtotrend = case_when(
time_frame %in% c("0:00 to 0:59", "1:00 to 1:59", "2:00 to 2:59", "3:00 to 3:59", "4:00 to 4:59", "5:00 to 5:59", "6:00 to 6:59", "7:00 to 7:59") ~ "0:00 to 7:59",

time_frame %in% c("8:00 to 8:59", "9:00 to 9:59", "10:00 to 10:59", "11:00 to 11:59", "12:00 to 12:59", "13:00 to 13:59", "14:00 to 14:59", "15:00 to 15:59") ~ "8:00 to 15:59",

time_frame %in% c("16:00 to 16:59", "17:00 to 17:59", "18:00 to 18:59", "19:00 to 19:59", "20:00 to 20:59","21:00 to 21:59", "22:00 to 22:59", "23:00 to 23:59") ~ "16:00 to 23:59",
    TRUE ~ NA_character_
  ))
```

#2 Time needed for a video to become trending (how many days needed for trending videos)

```
youtube$timetotrend <- as.Date(youtube$trending_date, format = "%y.%d.%m") - as.Date(youtube$publish_date, format = "%d/%m/%Y")

youtube$timetotrend <- as.factor(ifelse(youtube$timetotrend <= 7, youtube$timetotrend, "8+"))
```

2. Data visualization

2.1. General visualization

1. Trending videos by category and Trending videos by Country

#1 Distribution of trending videos by Categories

```
yt1<- data.frame(table(youtube$category_id))

ggplot(yt1, aes(x=reorder(Var1, -Freq),
y=Freq)) +
  geom_segment( aes(x=reorder(Var1, Freq),
xend=reorder(Var1, Freq), y=0, yend=Freq),
color="blue") +
  geom_point( color="blue", size=4, alpha=0.6)
+ theme_light() +coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()) +
  labs(title = "Figure 1: Distribution of trending
videos by Categories",
       caption = "Source : YouTube Trending
Videos Dataset",
       x = "Category", y = "Number of Videos")
```

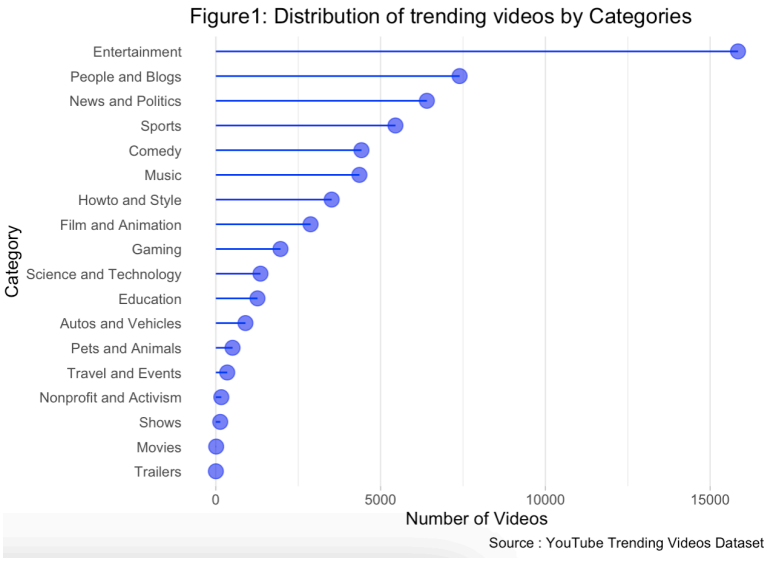


Figure1 shows that "Entertainment" has the most videos, more than double "People and blogs" coming in second, and "News and politics" ranked third. The 3 categories with the least number of videos are "Trailers", "Movies", "Shows".

#2 Distribution of trending videos by Country

```
yt2<-
data.frame(table(youtube$publish_country))

ggplot(yt2, aes(x=reorder(Var1, -Freq),
y=Freq)) +
geom_bar(stat="identity",fill="skyblue2",
color="grey") +theme_light() +coord_flip() +
  theme(panel.grid.major.y = element_blank(),
     panel.border = element_blank(),
  axis.ticks.y = element_blank()) + labs(title =
"Figure 2: Distribution of trending videos by
Country", caption = "Source: YouTube Trending
Videos Dataset", x = "Country", y = "Number of
Videos")
```

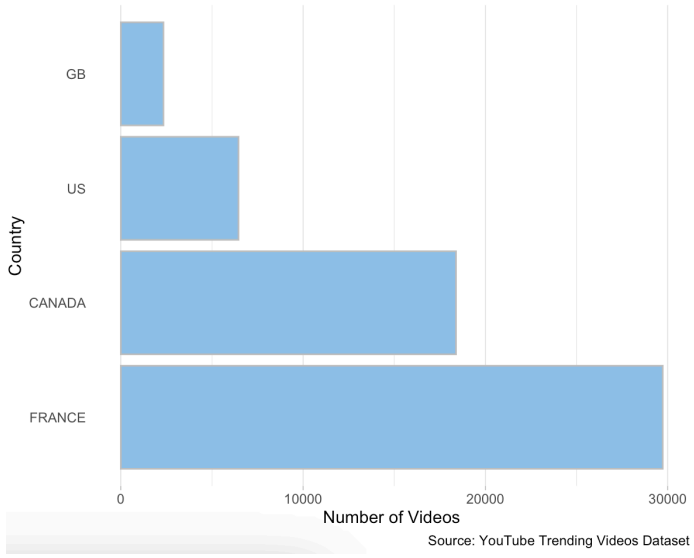
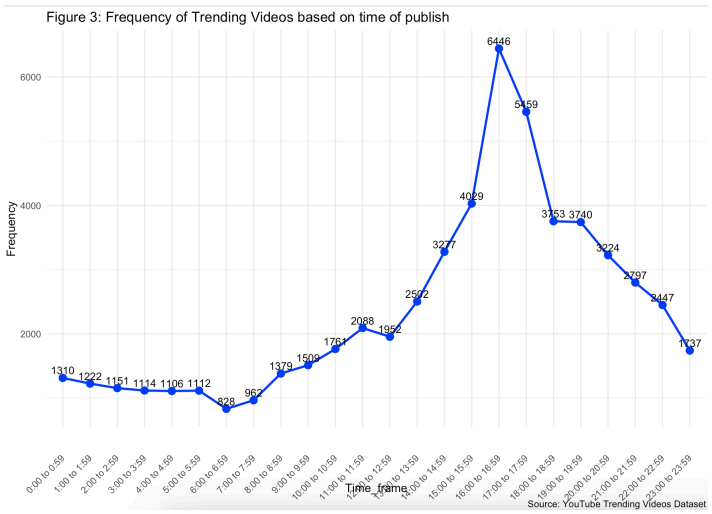


Figure 2 shows France is the country with the highest number of videos among the four countries, Canada is ranked number 2, then the US is number 3. GB is the country with the least number of videos.

2. Total trending videos based on: time of publish, period of publish

#3 Frequency of Trending Videos based on time of publish

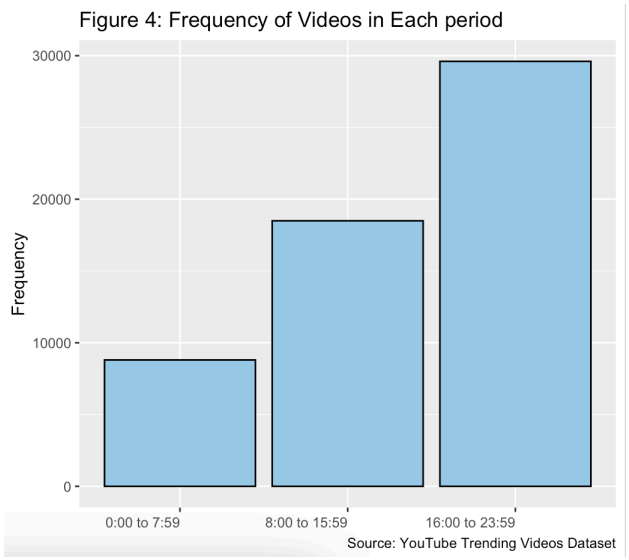
```
yt3 <- data.frame(table(factor(youtube$time_frame, levels = c("0:00 to 0:59", "1:00 to 1:59", "2:00 to 2:59", "3:00 to 3:59", "4:00 to 4:59", "5:00 to 5:59", "6:00 to 6:59", "7:00 to 7:59", "8:00 to 8:59", "9:00 to 9:59", "10:00 to 10:59", "11:00 to 11:59", "12:00 to 12:59", "13:00 to 13:59", "14:00 to 14:59", "15:00 to 15:59", "16:00 to 16:59", "17:00 to 17:59", "18:00 to 18:59", "19:00 to 19:59", "20:00 to 20:59","21:00 to 21:59", "22:00 to 22:59", "23:00 to 23:59"))))

ggplot(yt3, aes(x = Var1, y = Freq, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue", size = 3) +
  geom_text(aes(label = Freq), vjust = -0.5, size = 3.5, color = "black") +
  labs(title = "Figure 3: Frequency of Trending Videos based on time of publish",x = "Time_frame", y = "Frequency",caption = "Source: YouTube Trending Videos Dataset") + theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 1))
```



The time when the highest number of videos are posted is from 16:00. to 16:59 (6446 videos).
The time when the least number of videos are posted is from 6:00 to 6:59 (828 videos).
The number of uploaded videos gradually increased from 12:00 to 16:59 and then decreased until 11:59 p.m.

#4 Frequency of Videos in each period

```
youtube$periodtotrend <- factor(youtube$periodtotrend, levels = c('0:00 to 7:59','8:00 to 15:59','16:00 to 23:59'))

ggplot(youtube, aes(x = periodtotrend)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Figure 4: Frequency of Videos in Each period", x = "Category",y = "Frequency",
caption = "Source: YouTube Trending Videos Dataset") + theme(axis.text.x = element_text(vjust = 0.5, hjust = 1), axis.title.x = element_blank())
```

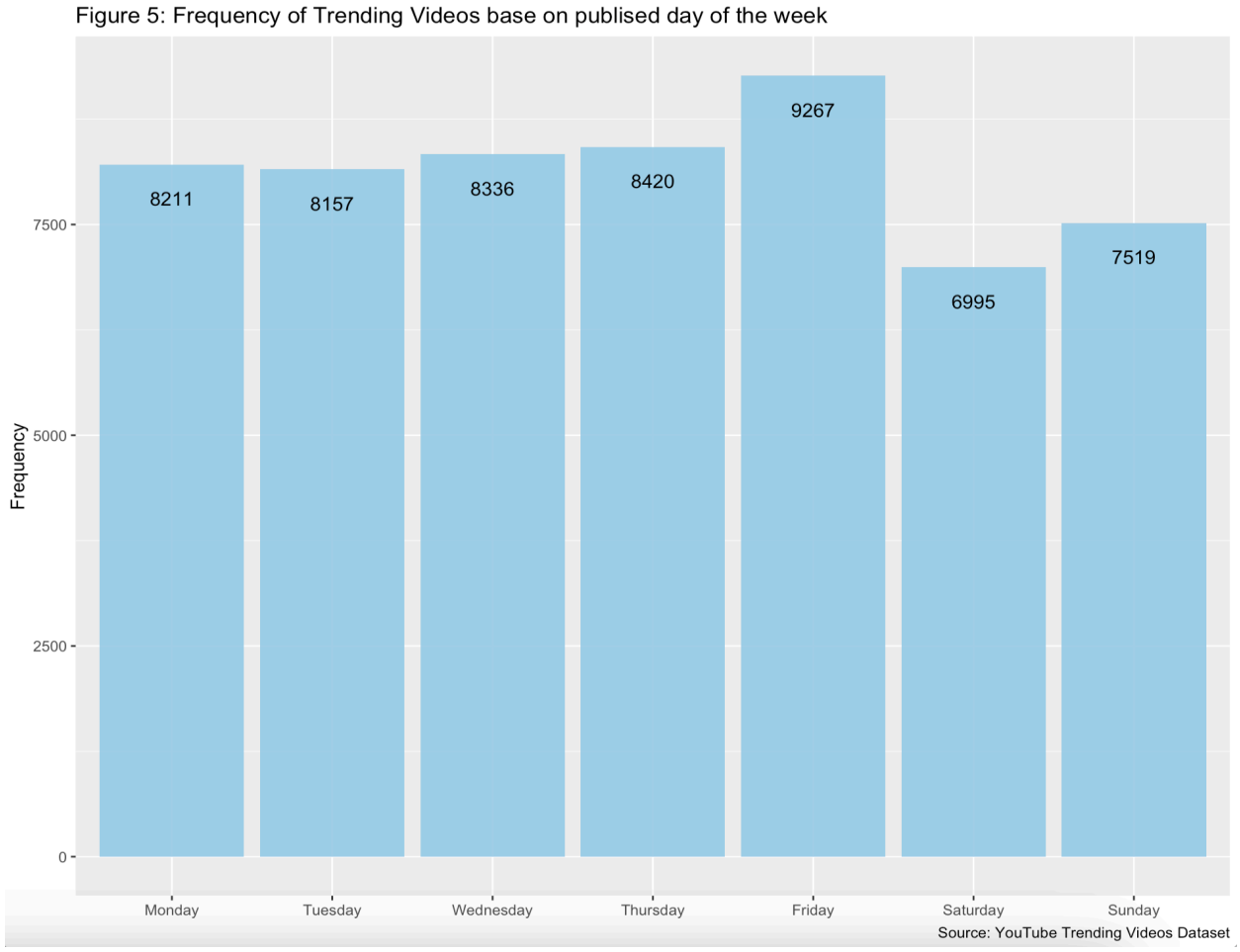


The first 8 hours of a day have the least amount of videos and then increase for the next 8 hours. The last 8 hours of the day from 16:00 p.m. to 23:59 p.m. have the highest number of videos posted.

3. Total trending videos based on publised day of the week

#5 Frequency of Trending Videos base on publised day of the week

```
yt5 <- data.frame(table(youtube$published_day_of_week))

yt5$Var1 <- factor(yt5$Var1, levels = c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday'))

ggplot(yt5, aes(x=Var1, y=Freq)) +
  geom_col(fill = "skyblue", alpha = 0.9) +
  labs(title = "Figure 5: Frequency of Trending Videos base on publised day of the week",
       x = NULL, y = "Frequency", caption = "Source: YouTube Trending Videos Dataset") +
  geom_text(aes(label = Freq), size = 4, hjust = 0.5, vjust = 3, position = "stack")
```



This bar chart shows that most of the videos were uploaded on Friday, while the least was on Saturday and then Sunday. The total number of videos posted on the remaining days of the week is quite even.

2.2. Specific visualization

```
#6 Chart of total views, likes, dislikes, comments of categories
library(dplyr)
library(gridExtra)
# Function to create plots
create_plot <- function(data, y_var, y_lab,fill_color) {
  yt6 <- data %>% select(category_id, {{y_var}}) %>% group_by(category_id) %>%
                                summarise({{y_var}} := sum({{y_var}})/1000000) %>% arrange(-{{y_var}})
  plot <- ggplot(yt6, aes(x = reorder(category_id, -{{y_var}}), y = {{y_var}})) +
    geom_bar(stat = "identity",fill=fill_color) + coord_flip() + theme_minimal() + labs(x = "Categories", y = y_lab)
  return(plot) }
# Create individual plots
plot1 <- create_plot(youtube, views, "Number of Million Views",c("skyblue3"))
plot2 <- create_plot(youtube, likes, "Number of Million Likes",c("green3"))
plot3 <- create_plot(youtube, dislikes, "Number of Million Dislikes",c("purple3"))
plot4 <- create_plot(youtube, comment_count, "Number of Million Comments",c("orange3"))
combined_plots <- grid.arrange(plot1, plot2, plot3, plot4, nrow = 2, ncol = 2)
show(combined_plots)
```

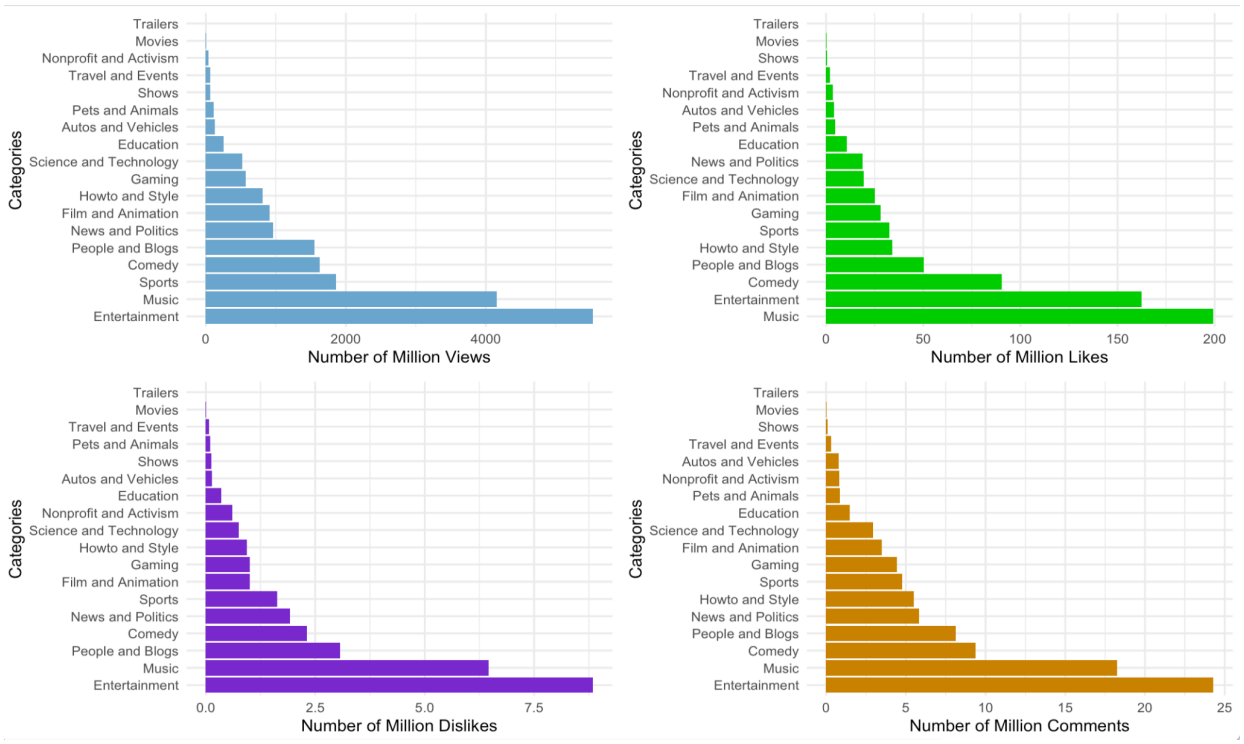


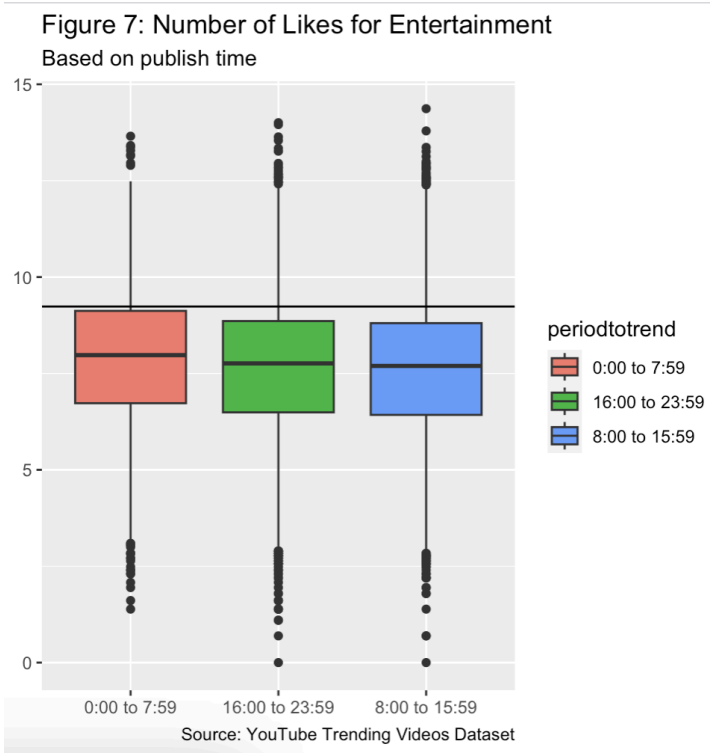
Figure 6: Chart of total views, likes, dislikes, comments of categories

"Entertainment" always occupies the highest position in terms of total views, total dislikes and total comments, probably partly because as shown in figure 1, "Entertainment" accounts for the largest number of videos. Meanwhile, "Music" only ranked 6th in terms of total number of videos however "Music" occupies the highest position about total likes, proving that music is loved by the audience and it can be said that the quality of music is better than entertainment.

#7 Number of Likes for Entertainment

#Entertainment is a category that always ranks 1 in terms of number of trending videos, views, dislikes or comments. So let's find out the distribution and average of likes based on the time of publication.

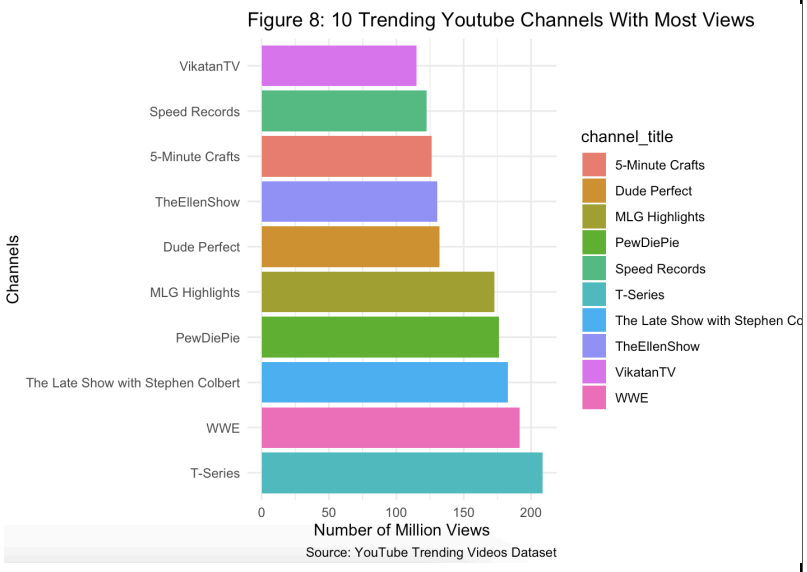
```
yt7 <- youtube %>%
  filter(category_id == "Entertainment") %>%
  select(periodtotrend, likes) %>%
  group_by(periodtotrend)

ggplot(data = yt7, aes(x = periodtotrend, y =
log(likes), fill = periodtotrend)) +
  geom_boxplot() + geom_hline(aes(yintercept
= log(mean(likes)))) +
  labs(title = "Number of Likes for
Entertainment", subtitle = "Based on publish
time ", caption = "Source: YouTube Trending
Videos Dataset", x = NULL, y = NULL)
```



0:00 to 7:59 is the best time to publish videos, due to highest average likes compared to other allocated times - Most data falls below average line.

#8: 10 Trending Youtube Channels With Most Views

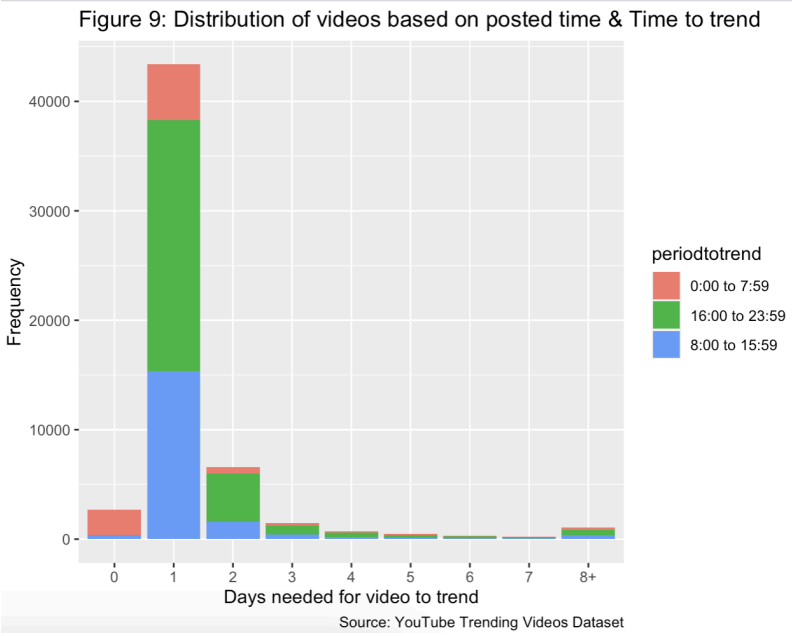
```
yt8 <- youtube %>%
  select(channel_title, views) %>%
  group_by(channel_title) %>%
  summarise(views = sum(views)/1000000)
%>% arrange(-views)

yt8_top10 <- data.frame(head(yt8, 10))

ggplot(yt8_top10, aes(x =reorder(channel_title,
-views), y = views, fill = channel_title)) +
  geom_bar(stat = "identity",position = "dodge")
+ coord_flip() + theme_minimal() + labs(title =
"Figure 8: 10 Trending Youtube Channels With
Most Views",caption = "Source: YouTube
Trending Videos Dataset", x = "Channels",
       y = "Number of Million Views")
```



“T-Series” is the channel which got the most views for their trending videos. Followed by “WWE” and “The Late Show with Stephen Colbert” ranked 2nd and 3rd respectively.

#9 Distribution of videos based on posted time & Time to trend

```
yt9 <-youtube %>%
  select(timetotrend, periodtotrend)

ggplot(data = yt9, aes(x = timetotrend)) +
  geom_bar(aes(fill = periodtotrend), position = "stack" ) +
  scale_x_discrete(guide = guide_axis(angle = 0)) +
  labs(title = "Figure 9: Distribution of videos based on posted time & Time to trend",
       caption = "Source: YouTube Trending Videos Dataset",
       x = 'Days needed for video to trend',
       y = 'Frequency')
```



The chart shows that the majority of videos only need 1 day for video to trend. The posting time for video to trend is from 16:00 to 23:59 and then the time frame from 8:00 to 15:59. Except for a small portion of videos that trend during the day when posted from 0:00 to 7:59.

#10: Correlation of view count, likes, dislikes, and comment count

```
library(corrplot)

columns_of_interest <- c("views", "likes", "dislikes", "comment_count")

correlation_matrix <- cor(youtube[columns_of_interest])
corrplot(correlation_matrix, method="pie")
```

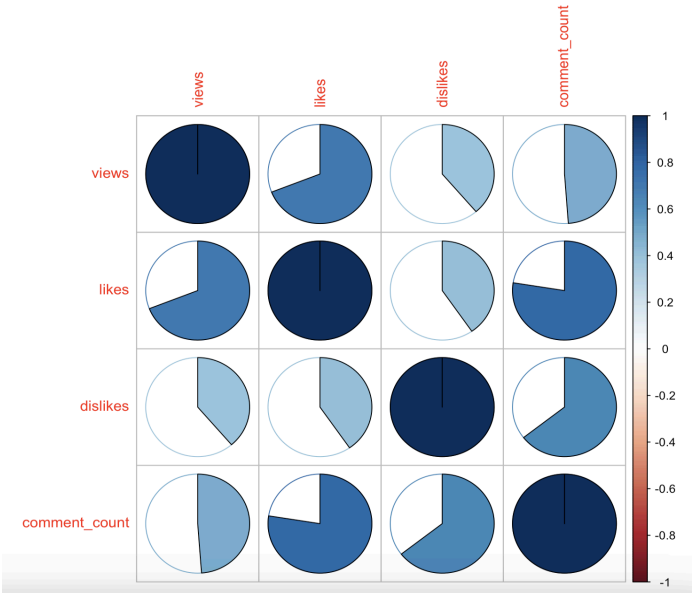


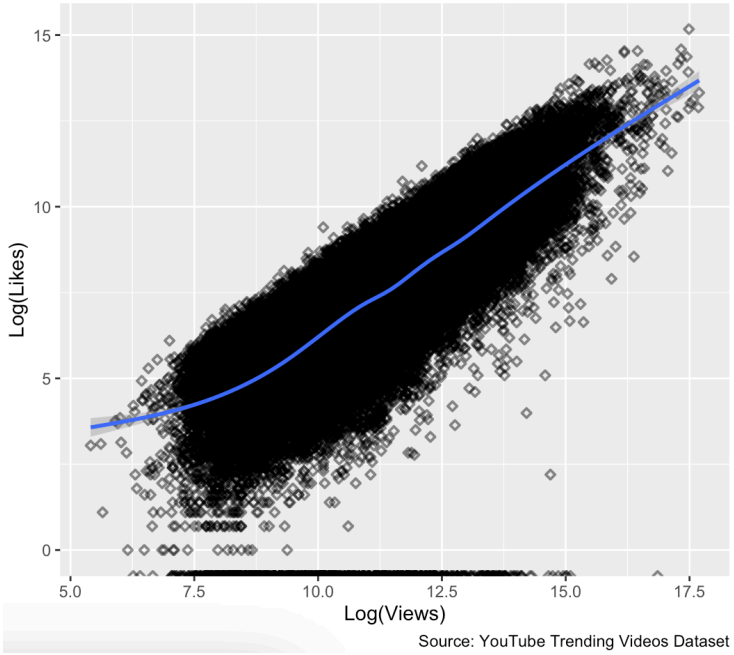
Figure 10: Correlation of view count, likes, dislikes, and comment count

The graph shows that there is a positive relation between views and likes, likes and comment_count, dislikes and comment_count.

#11: Correlation between Likes & Views

```
ggplot(data = youtube, aes(x = log(views), y = log(likes))) +
  geom_point(color = "black",
             fill = "#69b3a2",
             shape = 5,
             alpha = 0.5,
             size = 1,
             stroke = 1) +
  geom_smooth() +
  labs(title = "Figure 11: Correlation between Likes & Views",
       caption = "Source: YouTube Trending Videos Dataset",
       x = "Log(Views)",
       y = "Log(Likes)")
```

Figure 11: Correlation between Likes & Views

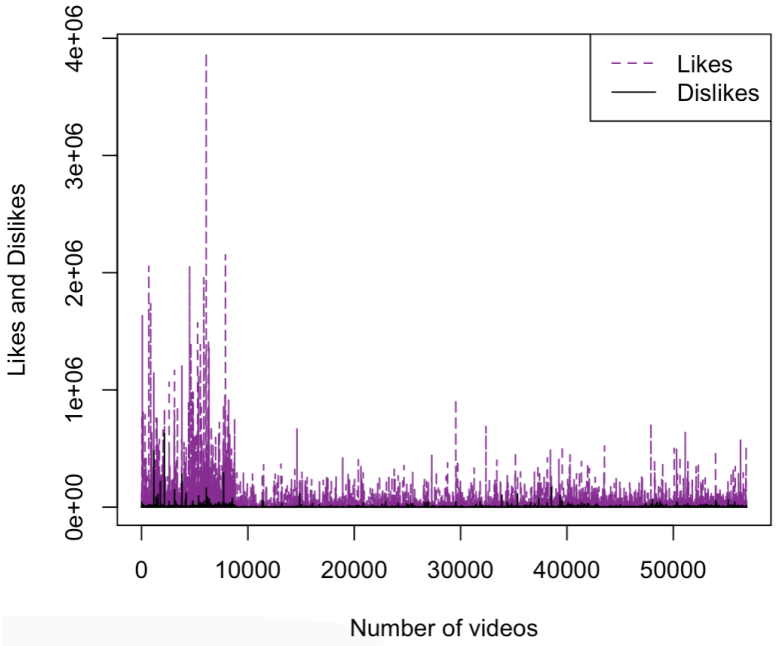


The chart shows that more views will lead to higher likes. The correlation between views and likes is a positive relation

#12 Correlation likes and dislikes

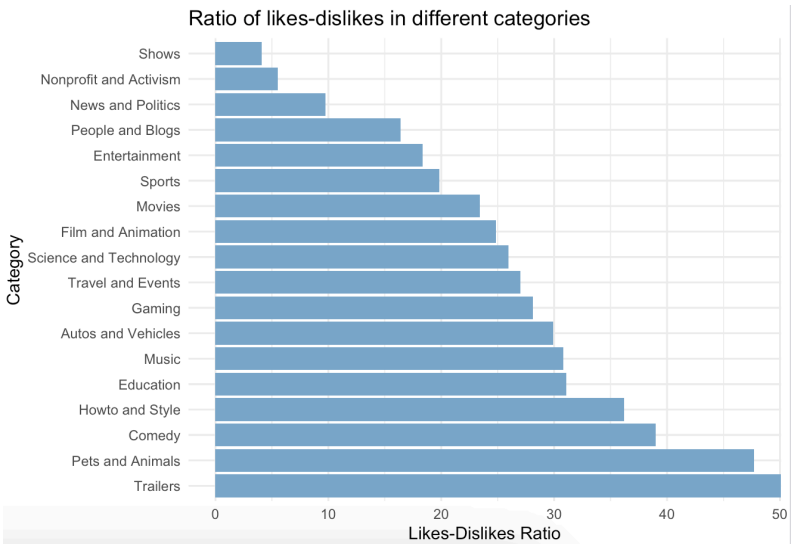
```
plot(youtube$likes, type = 'l', col = '#93089A', xlab = 'Number of videos', ylab = 'Likes and Dislikes',
     main = 'Figure 12: Correlation of Likes vs. Dislikes', lty = 2)

lines(youtube$dislikes, col = 'black')

legend('topright', legend = c('Likes', 'Dislikes'), col = c('#93089A', 'black'), lty = c(2, 1))
```

Figure 12: Correlation of Likes vs. Dislikes



The relationship between likes and dislikes across video counts. The purple line represents likes while the black line represents dislikes. It's evident from the plot that trending videos received significantly more likes than dislikes.

```r
#13 Calculating ratio likes-dislike for each category

likesdf <- youtube %>%
  group_by(category_id) %>%
  summarise(total_likes = sum(likes))

dislikesdf <- youtube %>%
  group_by(category_id) %>%
  summarise(total_dislikes = sum(dislikes))

# calculating ratios of likes to dislikes
ratiodf <- likesdf
ratiodf$ratio <- likesdf$total_likes /
dislikesdf$total_dislikes

# arranging categories by ratio, highest to lowest
ratiodf <- ratiodf[order(ratiodf$ratio, decreasing = TRUE), ]

# plotting bar chart
ggplot(ratiodf, aes(x = reorder(category_id, -ratio), y = ratio)) +
  geom_bar(stat = "identity", fill = "skyblue3") +
  labs(y = "Likes-Dislikes Ratio", x = "Category",title = "Ratio of likes-dislikes in different categories") +
  theme_minimal() +
  coord_flip()
```



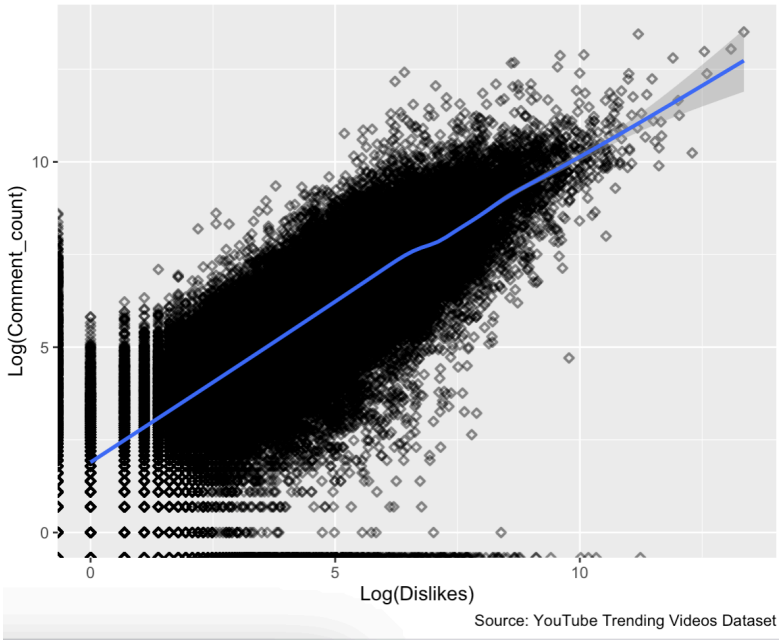
Videos about pets and animals, trailers get the most likes compared to dislikes among trending categories, while shows, nonprofit and activism,news and politics get the least.

This suggests people agree more on entertainment content, but news often splits opinions among users.

#14: Correlation between Dislikes & Comment_count

```
ggplot(data = youtube, aes(x = log(dislikes), y = log(comment_count))) +
  geom_point(color = "black",
             shape = 5,
             alpha = 0.5,
             size = 1,
             stroke = 1) +
  geom_smooth() +
  labs(title = "Figure 14: Correlation between Dislikes & Comment_count",
       caption = "Source: YouTube Trending Videos Dataset",
       x = "Log(Dislikes)",
       y = "Log(Comment_count)")
```

Figure 14: Correlation between Dislikes & Comment_count

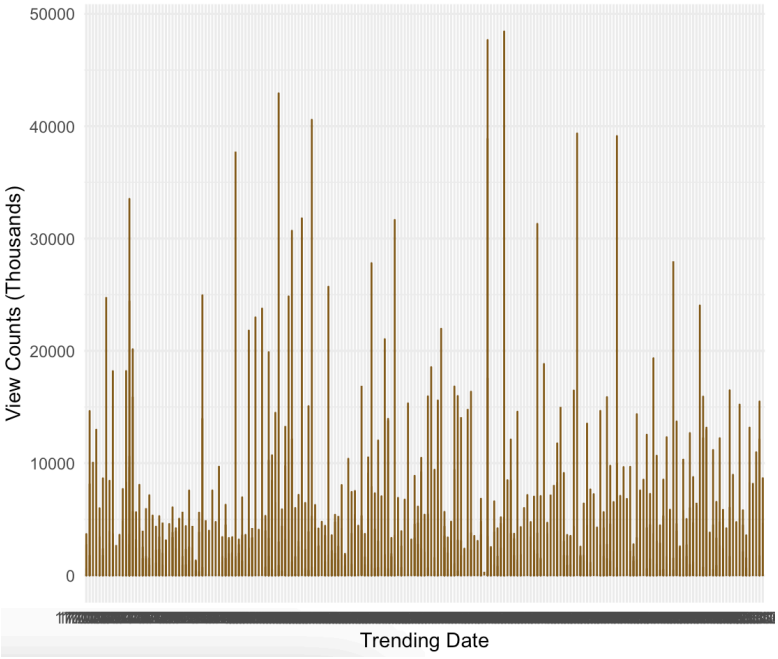


The correlation between dislikes and comment is a positive relation. The chart shows that more dislikes did lead to higher engagement rate.

#15: Views over time

```
ggplot(youtube, aes(x = trending_date, y = views/1000)) +
  geom_line(color = "orange4") +
  labs(title = "Figure 15: Views over time", x = "Trending Date", y = "View Counts (Thousands)") +
  theme_minimal()
```

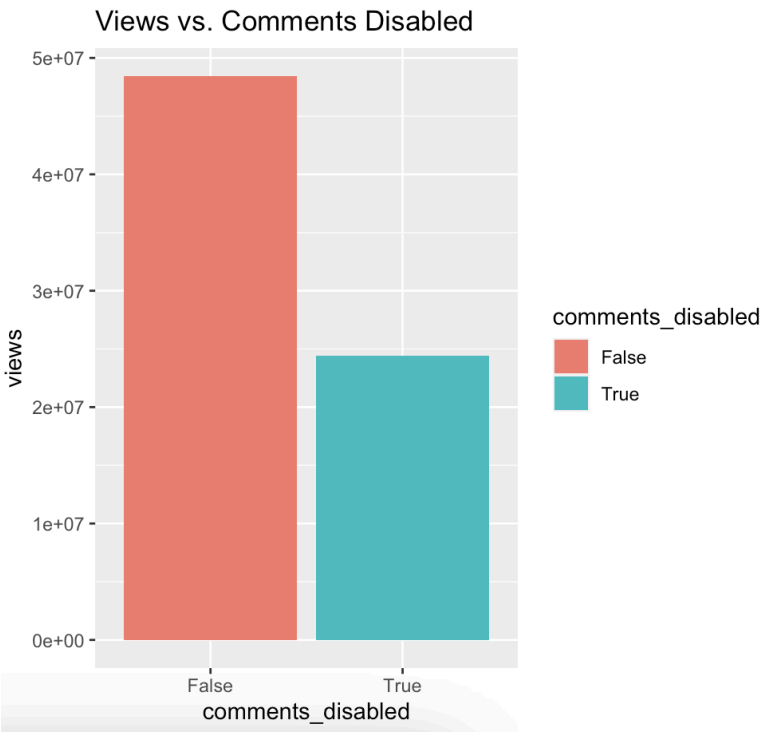
Figure 15: Views over time



The relationship between view counts and trending date. The highest total views in a day is up to nearly 50 million views.

#16: The relationship between view count and comments disabled.

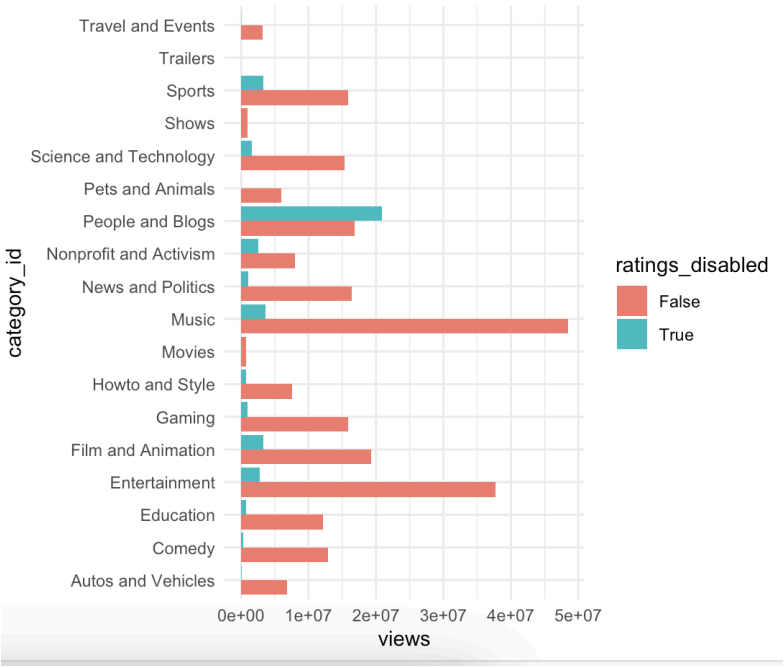
```
ggplot(youtube, aes(x = comments_disabled, y = views, fill = comments_disabled)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 16: Views vs. Comments Disabled")
```



As it is seen, comments disabled videos had fewer views than others.

#17 Relationship ratings_disabled with views per each category

```
ggplot(youtube, aes(x = views, y = category_id, fill = ratings_disabled)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal()
```

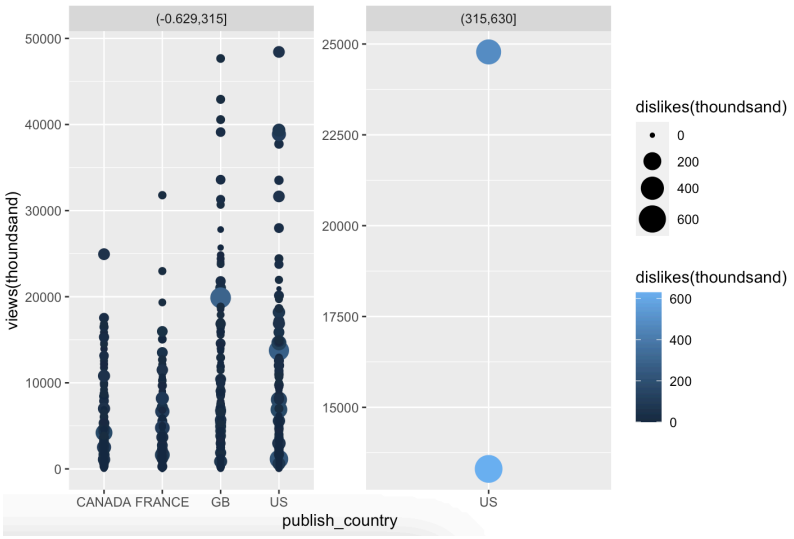


Ratings are mostly disabled for the “People and Blogs” category.
Ratings are not mostly disabled for the “Music” category.

#18 Relationship among Views, Country with Dislikes

```
ggplot(youtube, aes(x =
factor(publish_country), y = views/1000)) +
  geom_point(aes(size = dislikes/1000, color =
dislikes/1000)) +
  facet_wrap(~cut(dislikes/1000, breaks = 2),
scales = "free") +
  labs(x = "publish_country", y = "views", size =
"dislikes", color = "dislikes") +
  scale_size_continuous(range = c(1, 8)) +
  ggtitle("Figure 18: Relationship among Views,
Country with Dislikes")
```

Figure 18: Relationship among Views, Country with Dislikes

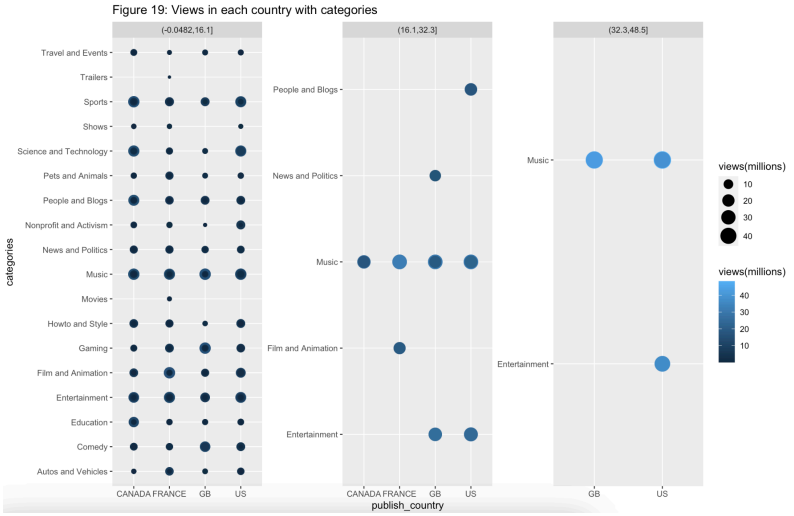


It can be seen that the number of views and dislikes in US and GB is higher than in the other two countries.
In term of US, the number of dislikes is more than 315,000 in both videos with few views and many views.

#19 Views in each country with categories

```
ggplot(youtube, aes(x =
factor(publish_country), y = category_id)) +
  geom_point(aes(size = views/1000000, color
= views/1000000)) +
  facet_wrap(~cut(views/1000000, breaks = 3),
scales = "free") +
  labs(x = "publish_country", y = "categories",
size = "views(millions)", color =
"views(millions)") +
  scale_size_continuous(range = c(1, 8)) +
  ggtitle("Figure 19: Views in each country with
categories")
```

Figure 19: Views in each country with categories



Range from 16.1 to 32.3 million views with only 5 categories "People and Blogs" in the US, "News and Politics in GB, "Music" in all 4 countries (of which France is the largest). "Film and Animation" in France, and "Entertainment" in GB vs US.
Range from 32.3 to 48.5 million views, only 2 countries GB and US for “Music”, and “Entertainment” only for US.

3. Conclusion

In summary, analyzing datasets across variables such as video categories, countries, posting times, and viewer engagement has led to the following conclusions:

- ❖ Entertainment" videos is always ranks 1 in total views, comments, and dislikes. Notably, "Music" stands out for its exceptional quality, having high likes despite a lower video count than “Entertainment”.
- ❖ Timing plays a crucial role, with 16:00 to 16:59 being the peak posting period, and Fridays witnessing the highest upload frequency, emphasizing the strategic importance of timing for optimal viewership and engagement.
- ❖ The correlation between dislikes and comments and the correlation between views and likes are a positive relation.
- ❖ It seems to be that people agree more on entertainment content, but news often splits opinions among users.
- ❖ About publishing countries, the US leading in views followed by Canada and France, GB.

4. Reference

- ➢ https://www.kaggle.com/datasets/thedevastator/youtube-trending-videos-dataset
- ➢ https://mixedanalytics.com/blog/list-of-youtube-video-category-ids/
- ➢ https://www.data-to-viz.com/