Youtube Streamer Analysis

Pythias

2024-02-06

Contents


```
setwd("C:/Users/Pythias/Desktop/Persoal Project/Intern Career/Task 1")
```

```
ysa=read.csv(file.choose())

library("janitor")

colnames(ysa)[4]="subscribers"

ysa=clean_names(ysa)
```

DATA EXPLORATION

Data Structure & Key Variables

```
library(dplyr)

str(ysa)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ username   : chr  "tseries" "MrBeast" "CoComelon" "SETIndia" ...
##  $ categories : chr  "Música y baile" "Videojuegos, Humor" "Educación" "" ...
##  $ subscribers: num  2.50e+08 1.84e+08 1.66e+08 1.63e+08 1.14e+08 ...
```

```
##  $ country    : chr  "India" "Estados Unidos" "Unknown" "India" ...
##  $ visits     : num  8.62e+04 1.17e+08 7.00e+06 1.56e+04 3.90e+06 ...
##  $ likes      : num  2700 5300000 24700 166 12400 ...
##  $ comments   : num  78 18500 0 9 0 4900 0 0 32 214 ...
##  $ links      : chr  "http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy4_brA" "http://youtube.com/channe
```

```
ysa_numeric = ysa %>%
  select(c(rank,subscribers,visits,likes,comments))

ysa_categorial = ysa %>%
  select(c(username,categories,country,links))
```

- The dataset has 4 character variables and 5 numerical variables
- The dataset has 1000 observations and 9 variables

Key Variables (Names)

```
names(ysa)
```

```
## [1] "rank"        "username"    "categories"  "subscribers" "country"
## [6] "visits"      "likes"       "comments"    "links"
```

The first 6 values of key variable names

```
head(ysa)
```

```
##   rank      username             categories subscribers        country
## 1    1       tseries         Música y baile   249500000          India
## 2    2       MrBeast     Videojuegos, Humor   183500000 Estados Unidos
## 3    3     CoComelon              Educación   165500000        Unknown
## 4    4      SETIndia                        162600000          India
## 5    5 KidsDianaShow    Animación, Juguetes   113500000        Unknown
## 6    6     PewDiePie Películas, Videojuegos   111500000 Estados Unidos
##      visits   likes comments
## 1     86200    2700       78
## 2 117400000 5300000    18500
## 3   7000000   24700        0
## 4     15600     166        9
## 5   3900000   12400        0
## 6   2400000  197300     4900
##                                            links
## 1 http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy4_brA
## 2 http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8uQQuVA
## 3 http://youtube.com/channel/UCbCmjCuTUZos6Inko4u57UQ
## 4 http://youtube.com/channel/UCpEhnqL0y41EpW2TvWAHD7Q
## 5 http://youtube.com/channel/UCk8GzjMOrta8yxDcKfylJYw
## 6 http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_AJ5Yw
```

Summary Stats for numeric variables

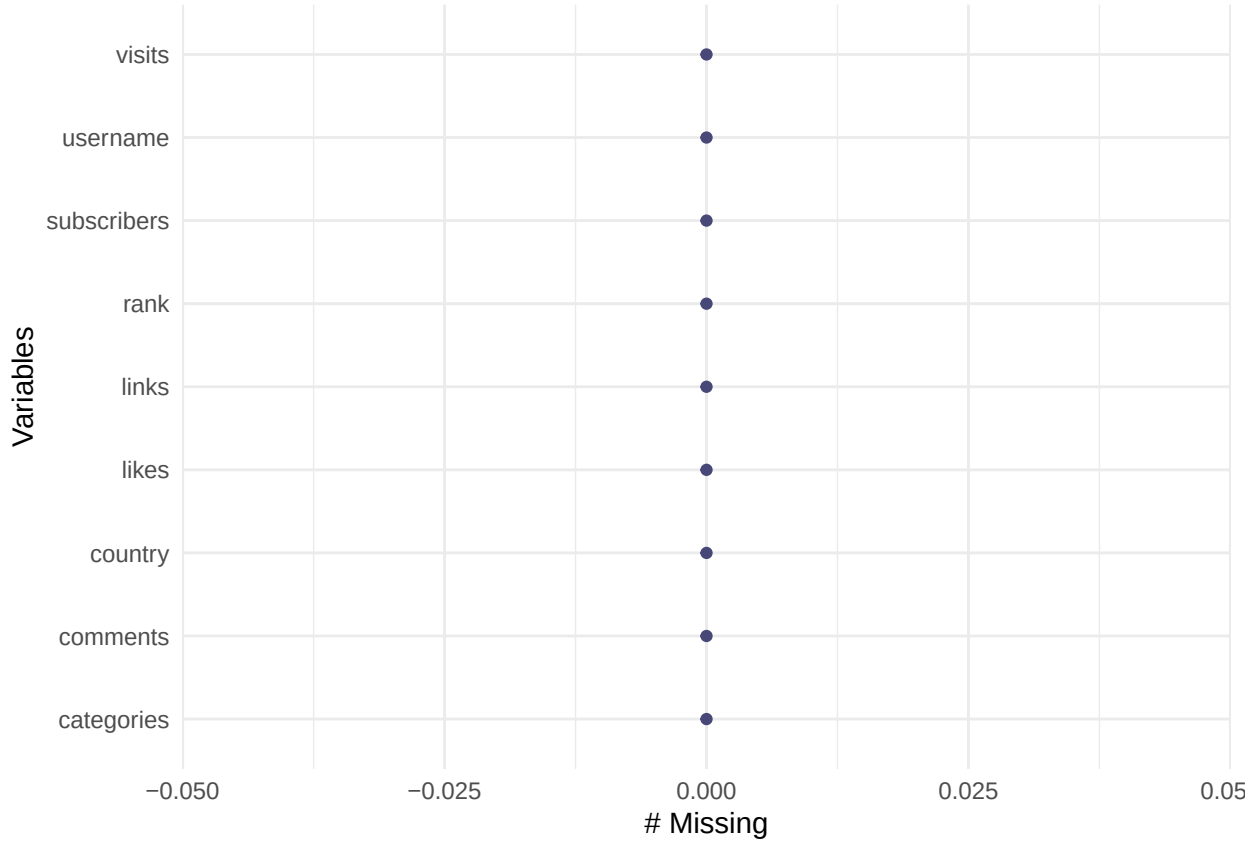
```
library(psych)
library(dplyr)
library(knitr)
ysa_numeric %>%
  summary() %>%
  kable()
```

rank	subscribers	visits	likes	comments
Min. : 1.0	Min. : 11700000	Min. : 0	Min. : 0	Min. : 0
1st Qu.: 250.8	1st Qu.: 13800000	1st Qu.: 31975	1st Qu.: 472	1st Qu.: 2
Median : 500.5	Median : 16750000	Median : 174450	Median : 3500	Median : 67
Mean : 500.5	Mean : 21894400	Mean : 1209446	Mean : 53633	Mean : 1289
3rd Qu.: 750.2	3rd Qu.: 23700000	3rd Qu.: 865475	3rd Qu.: 28650	3rd Qu.: 472
Max. :1000.0	Max. :249500000	Max. :117400000	Max. :5300000	Max. :154000

- Summary stats for each numeric variable

Missing Values

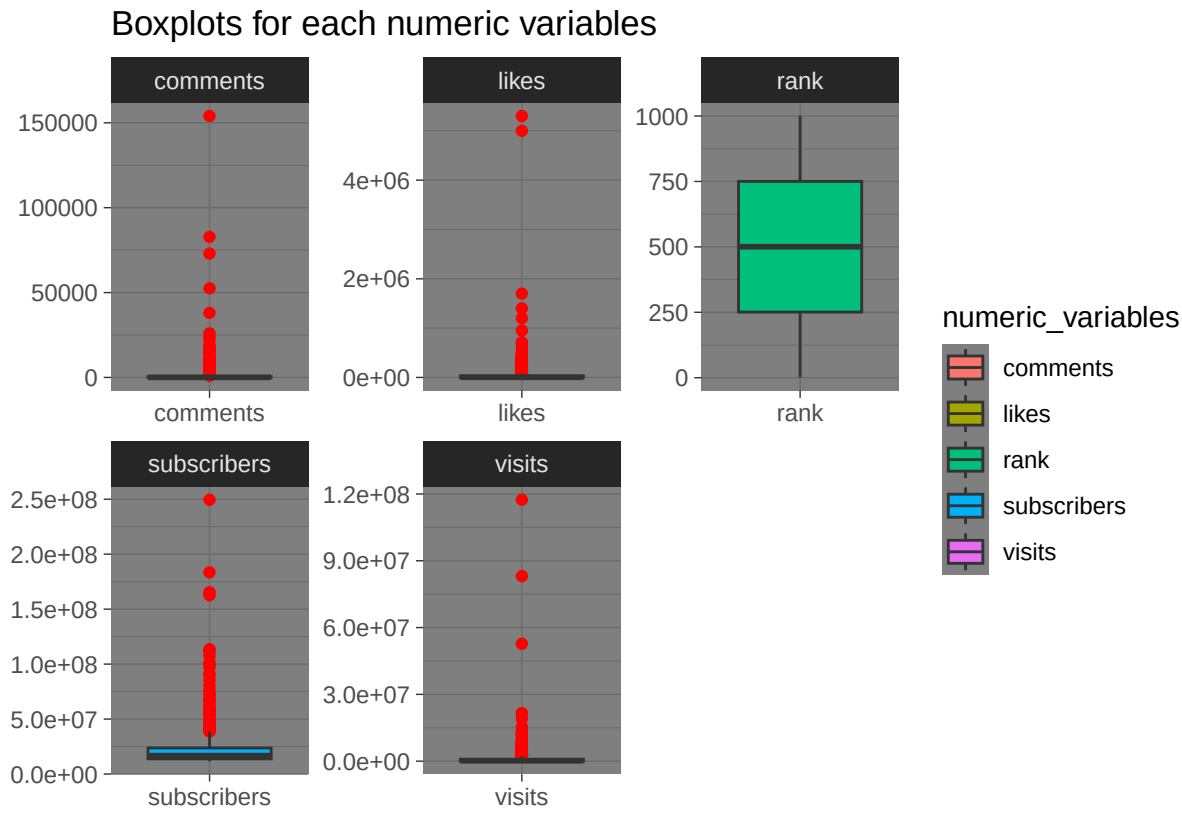
```
library(naniar)

ysa %>%
  gg_miss_var()
```



- The dataset has no missing values

Outliers

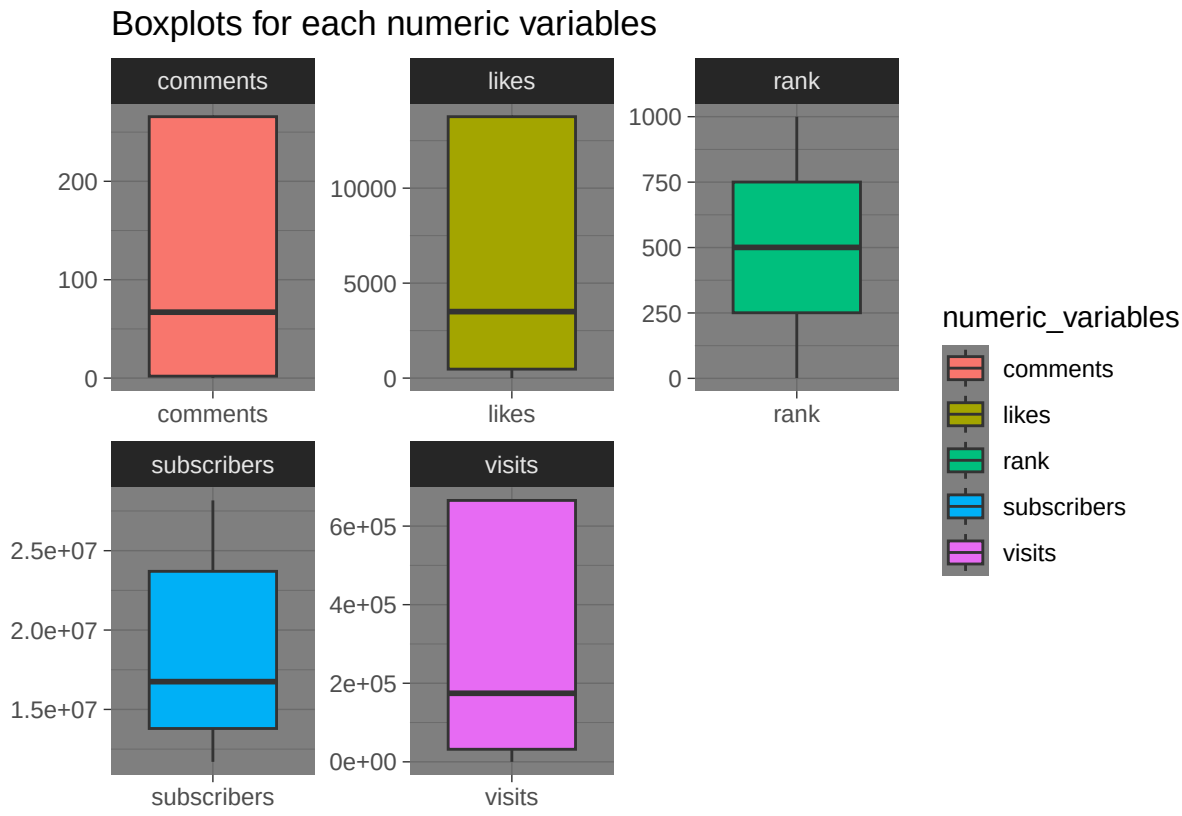
```
library(ggplot2)

library(tidyverse)

ysa_numeric_long = ysa_numeric %>%
  pivot_longer(everything(),
               names_to = "numeric_variables",
               values_to = "numeric_values")


ysa_numeric_long %>%
  ggplot(aes(numeric_variables,numeric_values))+
  geom_boxplot(aes(fill=numeric_variables),stat = "boxplot",position = "dodge",outlier.colour = "red")+
  facet_wrap(~ numeric_variables, scales = "free")+
  theme_dark()+labs(title = "Boxplots for each numeric variables",
                    x="",y="")
```



- the dataset contains outliers represented by the red circles for 4 numeric variables

#Handling outliers in the dataset

```r
library(robustHD)

ysa_numeric$subscribers=winsorize(ysa_numeric$subscribers,probs = c(0.05,0.95))
ysa_numeric$visits=winsorize(ysa_numeric$visits,probs = c(0.05,0.95))
ysa_numeric$likes=winsorize(ysa_numeric$likes,probs = c(0.05,0.95))
ysa_numeric$comments=winsorize(ysa_numeric$comments,probs = c(0.05,0.95))

#org dataset

ysa$subscribers=winsorize(ysa$subscribers,probs = c(0.05,0.95))
ysa$visits=winsorize(ysa$visits,probs = c(0.05,0.95))
ysa$likes=winsorize(ysa$likes,probs = c(0.05,0.95))
ysa$comments=winsorize(ysa$comments,probs = c(0.05,0.95))


ysa_numeric_long2 = ysa_numeric %>%
  pivot_longer(everything(),
               names_to = "numeric_variables",
               values_to = "numeric_values")


ysa_numeric_long2 %>%
  ggplot(aes(numeric_variables,numeric_values))+
  geom_boxplot(aes(fill=numeric_variables),stat = "boxplot",position = "dodge",outlier.colour = "red")+
  facet_wrap(~ numeric_variables, scales = "free")+
  theme_dark()+labs(title = "Boxplots for each numeric variables",
                    x="",y="")
```



- Handled outliers using robust method
- As shown by the box plots there are no longer outliers in the dataset

TREND ANALYSIS

Popular category

#Trends among the top YouTube streamers

```
library(knitr)
library(dplyr)


table(ysa$categories) %>%
  kable(caption = "Most popular categories")
```

Table 2: Most popular categories

Var1	Freq
	306
Animación	22
Animación, Humor	27

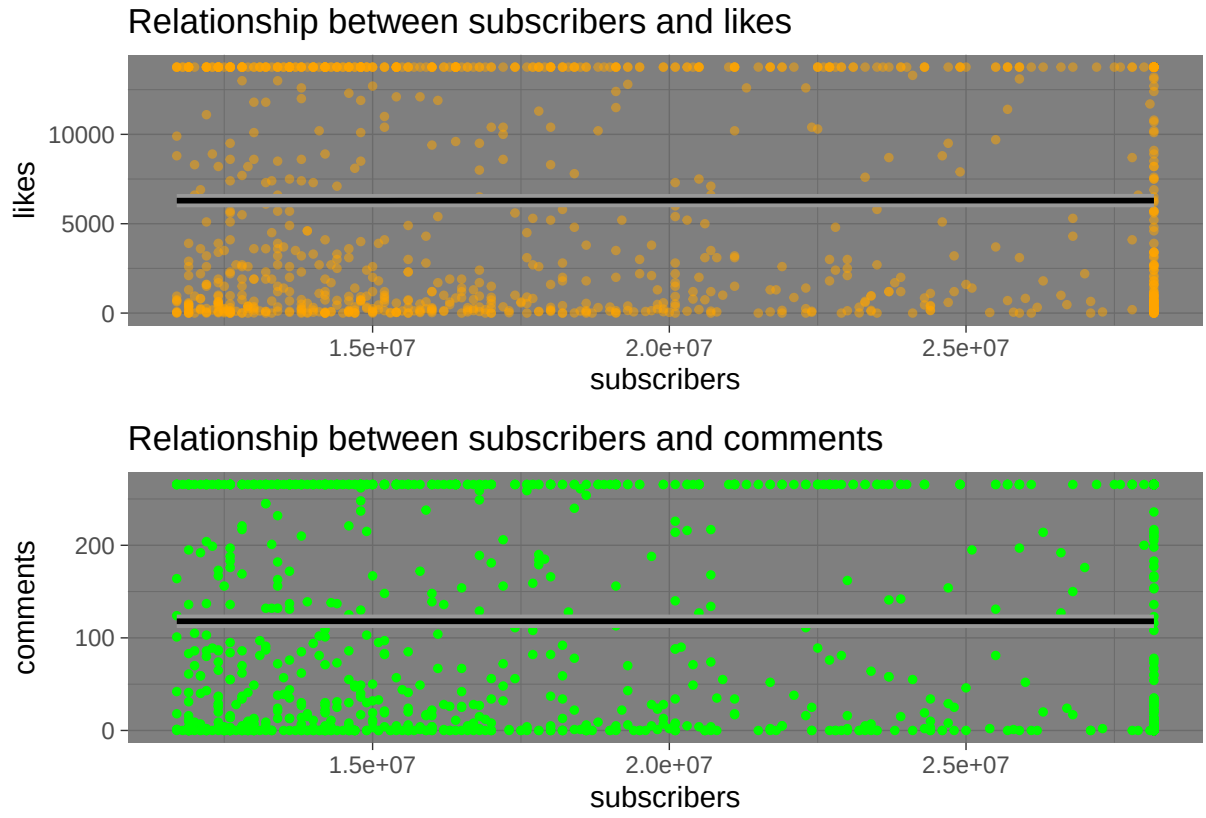
Var1	Freq
Animación, Humor, Juguetes	1
Animación, Juguetes	29
Animación, Videojuegos	34
Animales y mascotas	2
ASMR	1
ASMR, Comida y bebida	1
Belleza, Moda	1
Ciencia y tecnología	14
Coches y vehículos	2
Comida y bebida	12
Comida y bebida, Juguetes	1
Comida y bebida, Salud y autoayuda	1
Deportes	8
Diseño/arte	1
Diseño/arte, Belleza	1
Diseño/arte, DIY y Life Hacks	1
DIY y Life Hacks	3
DIY y Life Hacks, Juguetes	1
Educación	24
Educación, Juguetes	2
Fitness	2
Fitness, Salud y autoayuda	3
Humor	10
Juguetes	10
Juguetes, Coches y vehículos	4
Juguetes, DIY y Life Hacks	1
Moda	2
Música y baile	160
Música y baile, Animación	16
Música y baile, Humor	6
Música y baile, Juguetes	1
Música y baile, Películas	41
Noticias y Política	36
Películas	24
Películas, Animación	61
Películas, Humor	34
Películas, Juguetes	9
Películas, Videojuegos	8
Viajes, Espectáculos	1
Videojuegos	19
Videojuegos, Humor	17
Videojuegos, Juguetes	3
Vlogs diarios	37

- Categories with unknown names are the most popular with a record of 306.
- Música y baile is the second popular category with a frequency of 160.

Correlation

#Correlation between the number of subscribers and the number of likes or comments

```
sc1=ggplot(ysa,aes(subscribers,likes))+geom_point(color="orange",alpha=0.5,
                                                  size=1)+
  geom_smooth(color="black",alpha=1)+labs(
    title = "Relationship between subscribers and likes")+theme_dark()


sc2=ggplot(ysa,aes(subscribers,comments))+geom_point(color="green",alpha=1,
                                                  size=1)+
  geom_smooth(color="black",alpha=1)+labs(
    title = "Relationship between subscribers and comments")+theme_dark()

library(patchwork)

sc1 / sc2
```




```
new_cor=cor(ysa_numeric)

kable(new_cor,caption = "correlations")
```

Table 3: correlations

	rank	subscribers	visits	likes	comments
rank	1.0000000	-0.9653892	-0.0935175	-0.0266714	0.0223367
subscribers	-0.9653892	1.0000000	0.0946686	0.0232043	-0.0280959

	rank	subscribers	visits	likes	comments
visits	-0.0935175	0.0946686	1.0000000	0.8173862	0.6546486
likes	-0.0266714	0.0232043	0.8173862	1.0000000	0.8154030
comments	0.0223367	-0.0280959	0.6546486	0.8154030	1.0000000

- Visits and likes have a strong positive relationship (r=0.82) whilst subscribers and likes have a weak positive relationship

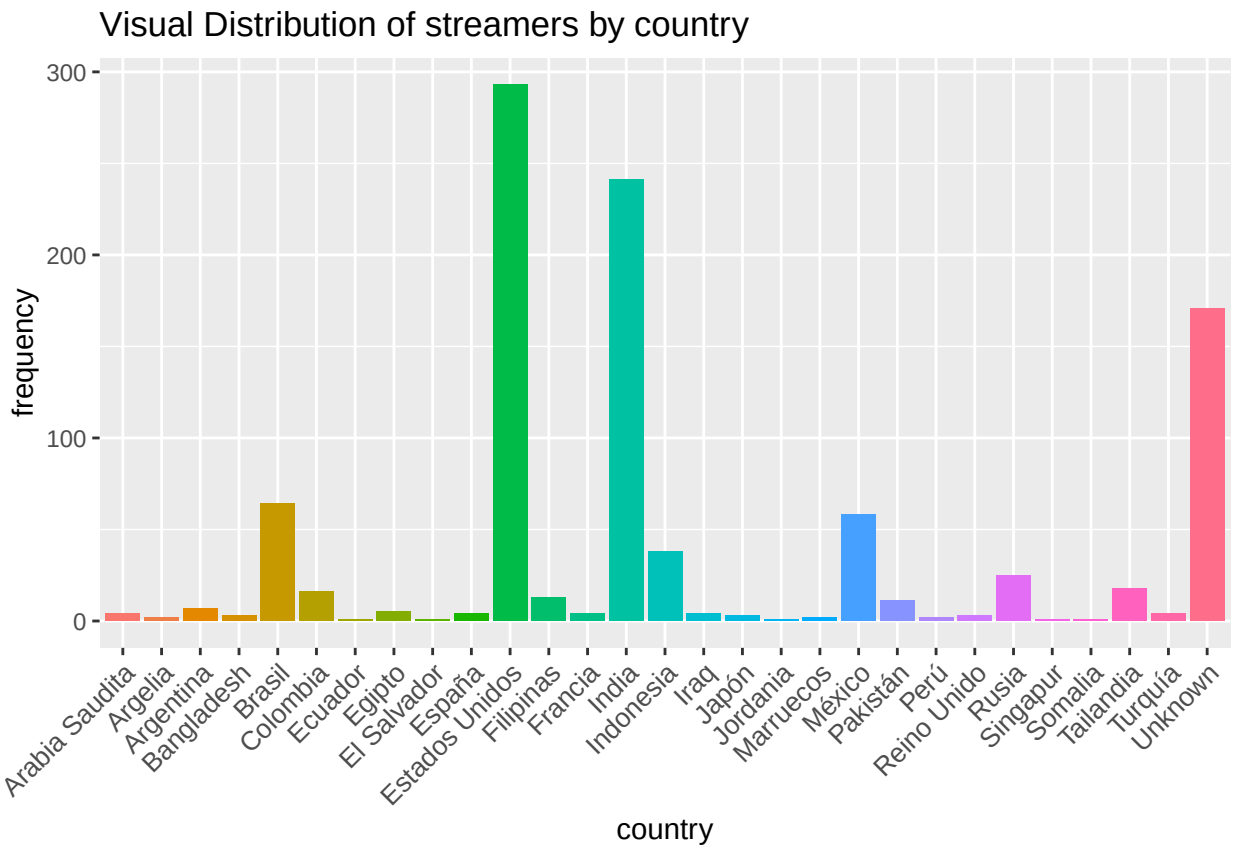
AUDIENCE STUDY

Distribution of streamers audiences by country

```
country_counts=table(ysa$country)
country_counts=as.data.frame(country_counts)
colnames(country_counts)=c("country","frequency")
#sorting
country_counts=country_counts[order(-country_counts$frequency),]

kable(country_counts)
```

	country	frequency
11	Estados Unidos	293
14	India	241
29	Unknown	171
5	Brasil	64
20	México	58
15	Indonesia	38
24	Rusia	25
27	Tailandia	18
6	Colombia	16
12	Filipinas	13
21	Pakistán	11
3	Argentina	7
8	Egipto	5
1	Arabia Saudita	4
10	España	4
13	Francia	4
16	Iraq	4
28	Turquía	4
4	Bangladesh	3
17	Japón	3
23	Reino Unido	3
2	Argelia	2
19	Marruecos	2
22	Perú	2
7	Ecuador	1
9	El Salvador	1
18	Jordania	1
25	Singapur	1
26	Somalia	1

```
ggplot(country_counts,aes(country,frequency,fill=country))+
  geom_bar(stat = "identity",show.legend = F)+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 45))+
  theme(legend.position ="bottom")+labs(title = "Visual Distribution of streamers by country")
```



- Estados has the highest number of streamers (293 audiences) followed by India with 241 audiences.
- 171 audiences are from unknown countries

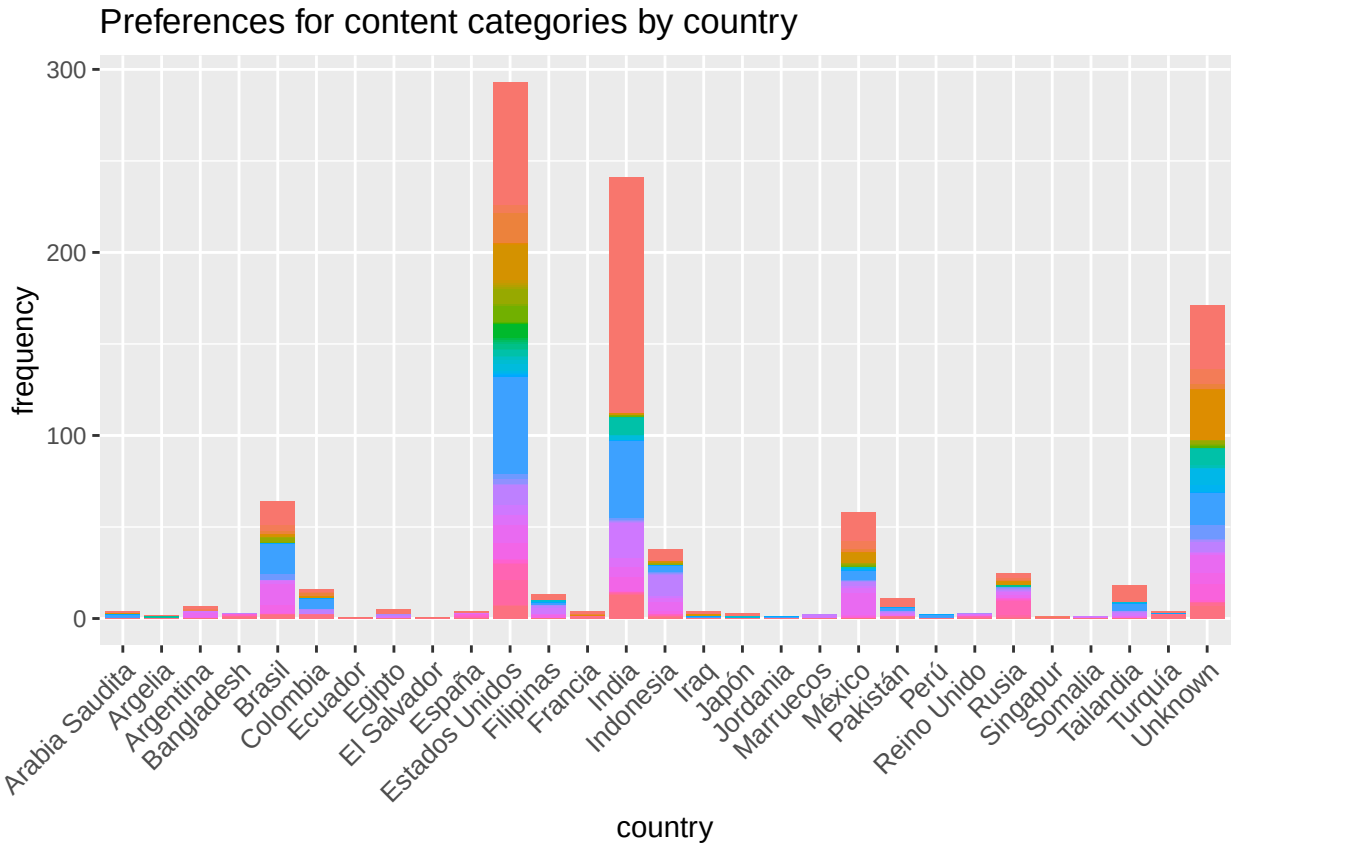
Regional preferences for specific content categories

```
library(knitr)

country_categories_count=table(ysa$country,ysa$categories)
country_categories_count=as.data.frame(country_categories_count)
colnames(country_categories_count)=c("country","categories","frequency")
#sorting

country_categories_count=country_categories_count[order(-country_categories_count$frequency),]

asss=kable(country_categories_count)

head(asss,11)
```

```
##  [1] "|     |country        |categories                        | frequency|"
```

```
##  [2] "|:----|:-------------|:----------------------------------|---------:|"
##  [3] "|14   |India         |                                   |       129|"
##  [4] "|11   |Estados Unidos |                                   |        67|"
##  [5] "|881  |Estados Unidos |Música y baile                     |        53|"
##  [6] "|884  |India         |Música y baile                     |        42|"
##  [7] "|29   |Unknown       |                                   |        35|"
##  [8] "|145  |Unknown       |Animación, Juguetes                |        28|"
##  [9] "|156  |Estados Unidos |Animación, Videojuegos             |        19|"
## [10] "|1029 |India         |Noticias y Política                |        19|"
## [11] "|899  |Unknown       |Música y baile                     |        18|"
```

```
ggplot(country_categories_count,aes(country,frequency,fill=categories))+
  geom_bar(stat = "identity",show.legend = F,position = "stack")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 45))+
  theme(legend.position ="bottom")+labs(title = "Preferences for content categories by country")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 45))
```



- there are regional preferences for specific content categories such as Musica y baile

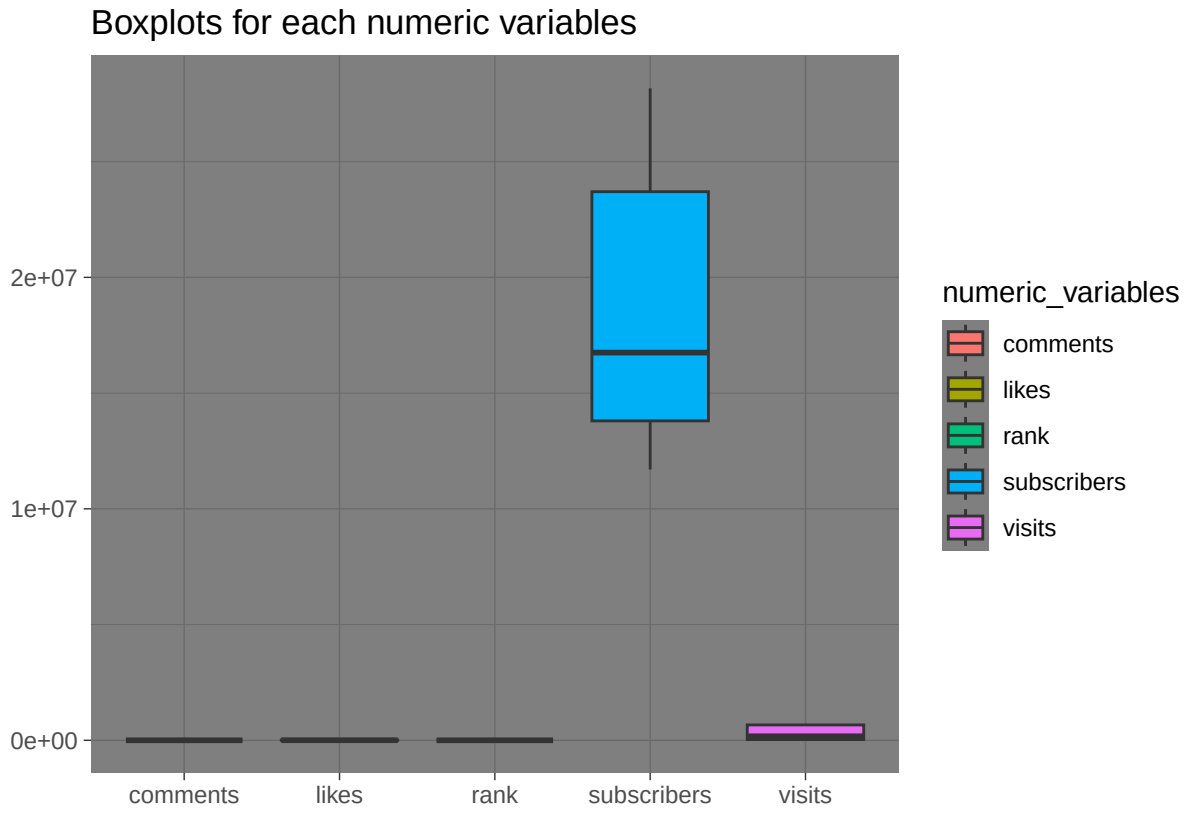
PERFORMANCE METRICS

The average number of subscribers, visits, likes, and comments.

```
stats=kable(summary(ysa_numeric))
stats
```

rank	subscribers	visits	likes	comments
Min. : 1.0	Min. :11700000	Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 250.8	1st Qu.:13800000	1st Qu.: 31975	1st Qu.: 471.8	1st Qu.: 2.0
Median : 500.5	Median :16750000	Median :174450	Median : 3500.0	Median : 67.0
Mean : 500.5	Mean :18709023	Mean :293544	Mean : 6292.1	Mean :117.9
3rd Qu.: 750.2	3rd Qu.:23700000	3rd Qu.:665339	3rd Qu.:13762.6	3rd Qu.:265.7
Max. :1000.0	Max. :28166020	Max. :665339	Max. :13762.6	Max. :265.7

```
ysa_numeric_long2 %>%
  ggplot(aes(numeric_variables,numeric_values))+
  geom_boxplot(aes(fill=numeric_variables),stat = "boxplot",position = "dodge",outlier.colour = "red")+
  theme_dark()+labs(title = "Boxplots for each numeric variables",
                    x="",y="")
```



- Subscribers have the highest average number

CONTENT CATEGORIES

```r
library(knitr)

cc=table(ysa$categories)

cc1=as.data.frame(cc)

cc1=cc1[order(-cc1$Freq),]

kable(cc1)
```

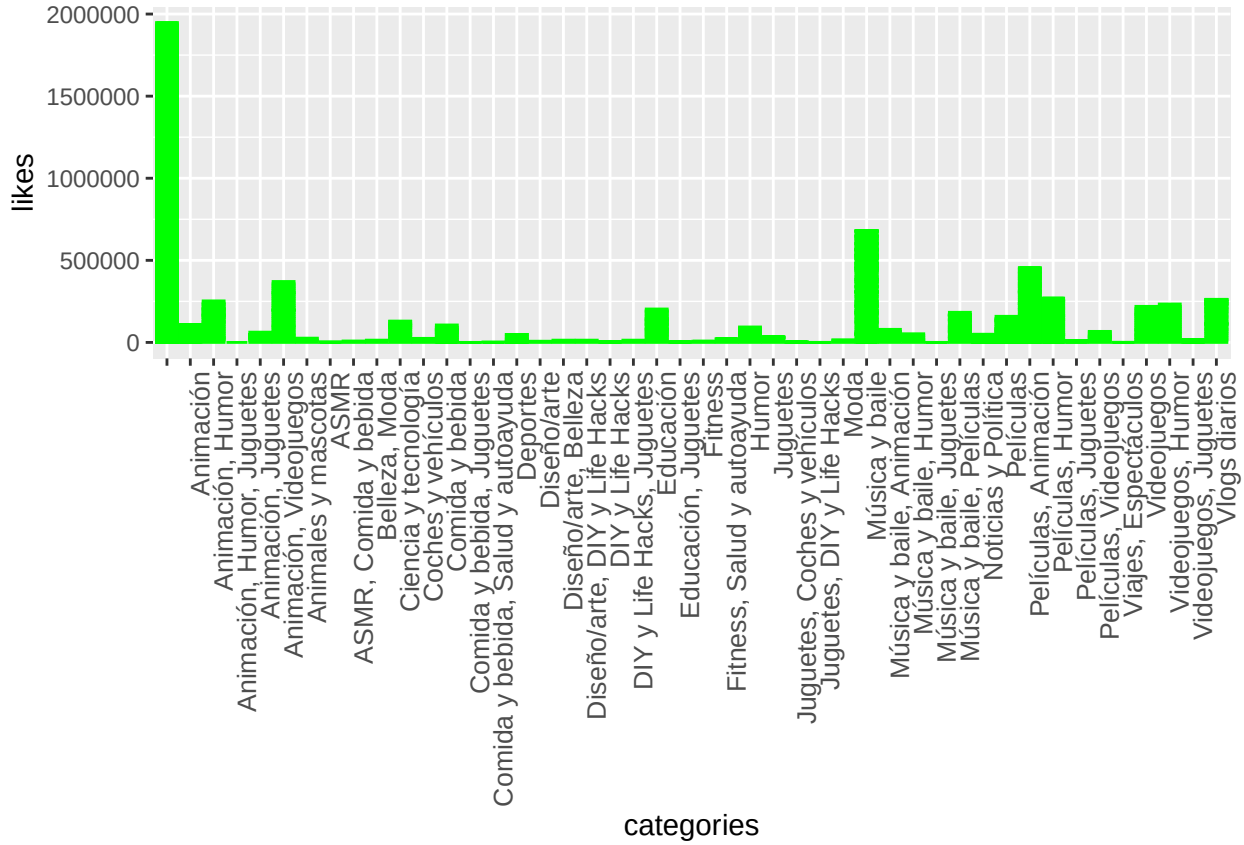
	Var1	Freq
1		306
31	Música y baile	160
38	Películas, Animación	61
35	Música y baile, Películas	41
46	Vlogs diarios	37
36	Noticias y Política	36
6	Animación, Videojuegos	34
39	Películas, Humor	34
5	Animación, Juguetes	29
3	Animación, Humor	27
22	Educación	24
37	Películas	24
2	Animación	22
43	Videojuegos	19
44	Videojuegos, Humor	17
32	Música y baile, Animación	16
11	Ciencia y tecnología	14
13	Comida y bebida	12
26	Humor	10
27	Juguetes	10
40	Películas, Juguetes	9
16	Deportes	8
41	Películas, Videojuegos	8
33	Música y baile, Humor	6
28	Juguetes, Coches y vehículos	4
20	DIY y Life Hacks	3
25	Fitness, Salud y autoayuda	3
45	Videojuegos, Juguetes	3
7	Animales y mascotas	2
12	Coches y vehículos	2
23	Educación, Juguetes	2
24	Fitness	2
30	Moda	2
4	Animación, Humor, Juguetes	1
8	ASMR	1
9	ASMR, Comida y bebida	1
10	Belleza, Moda	1
14	Comida y bebida, Juguetes	1
15	Comida y bebida, Salud y autoayuda	1
17	Diseño/arte	1
18	Diseño/arte, Belleza	1

	Var1	Freq
19	Diseño/arte, DIY y Life Hacks	1
21	DIY y Life Hacks, Juguetes	1
29	Juguetes, DIY y Life Hacks	1
34	Música y baile, Juguetes	1
42	Viajes, Espectáculos	1

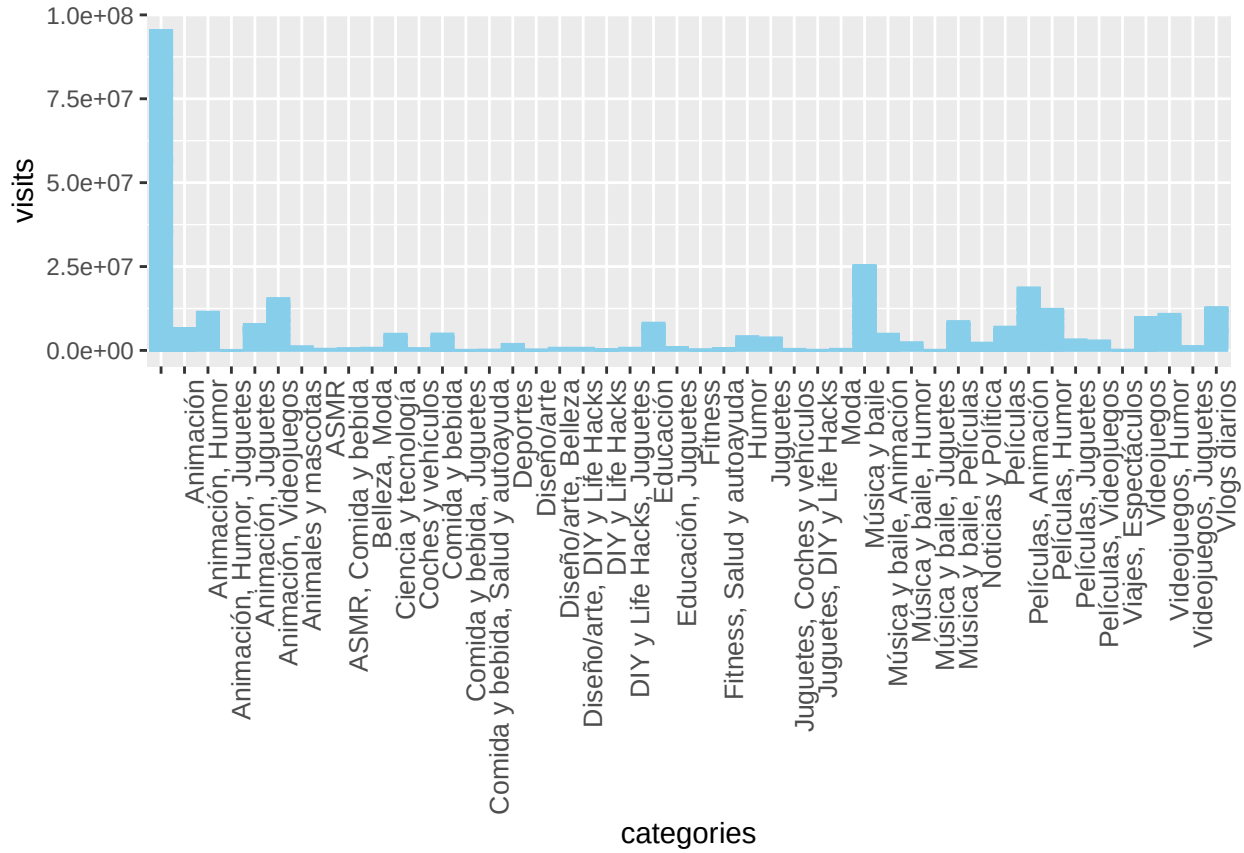
- Categories with highest number of streamers is unknown (306 streamers)

Categories with exceptional performance matrices

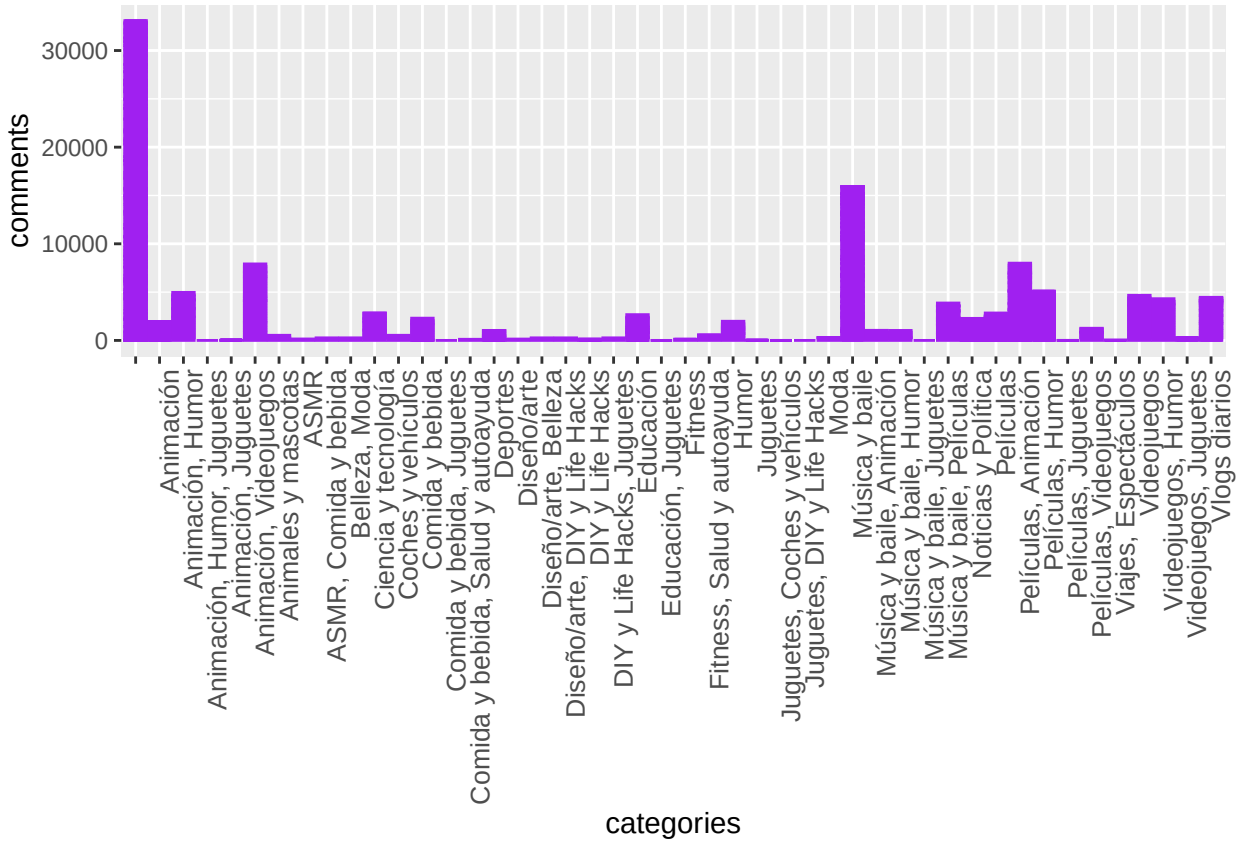
```
cxp= ysa %>%
  select(c("categories","likes","comments","subscribers","visits"))

pe1=ggplot(ysa, aes(categories,likes))+
  geom_bar(stat="identity",color="green")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 90))

pe2=ggplot(ysa, aes(categories,visits))+
  geom_bar(stat="identity",color="skyblue")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 90))


pe3=ggplot(ysa, aes(categories,comments))+
  geom_bar(stat="identity",color="purple")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 90))


pe4=ggplot(ysa, aes(categories,subscribers),color="black")+
  geom_bar(stat="identity",color="orange")+
  theme(axis.text.x = element_text(size = 10, hjust=1,angle = 90))


pe1
```


2000000
1500000
1000000
500000
0
likes
Animación
Animación, Humor
Animación, Humor, Juguetes
Animación, Juguetes
Animación, Videojuegos
Animales y mascotas
ASMR
ASMR, Comida y bebida
Belleza, Moda
Ciencia y tecnología
Coches y vehículos
Comida y bebida
Comida y bebida, Juguetes
Comida y bebida, Salud y autoayuda
Deportes
Diseño/arte
Diseño/arte, Belleza
Diseño/arte, DIY y Life Hacks
DIY y Life Hacks
DIY y Life Hacks, Juguetes
Educación
Educación, Juguetes
Fitness
Fitness, Salud y autoayuda
Humor
Juguetes
Juguetes, Coches y vehículos
Juguetes, DIY y Life Hacks
Moda
Música y baile
Música y baile, Animación
Música y baile, Humor
Música y baile, Juguetes
Música y baile, Películas
Noticias y Política
Películas
Películas, Animación
Películas, Humor
Películas, Juguetes
Películas, Videojuegos
Viajes, Espectáculos
Videojuegos
Videojuegos, Humor
Videojuegos, Juguetes
Vlogs diarios
categories

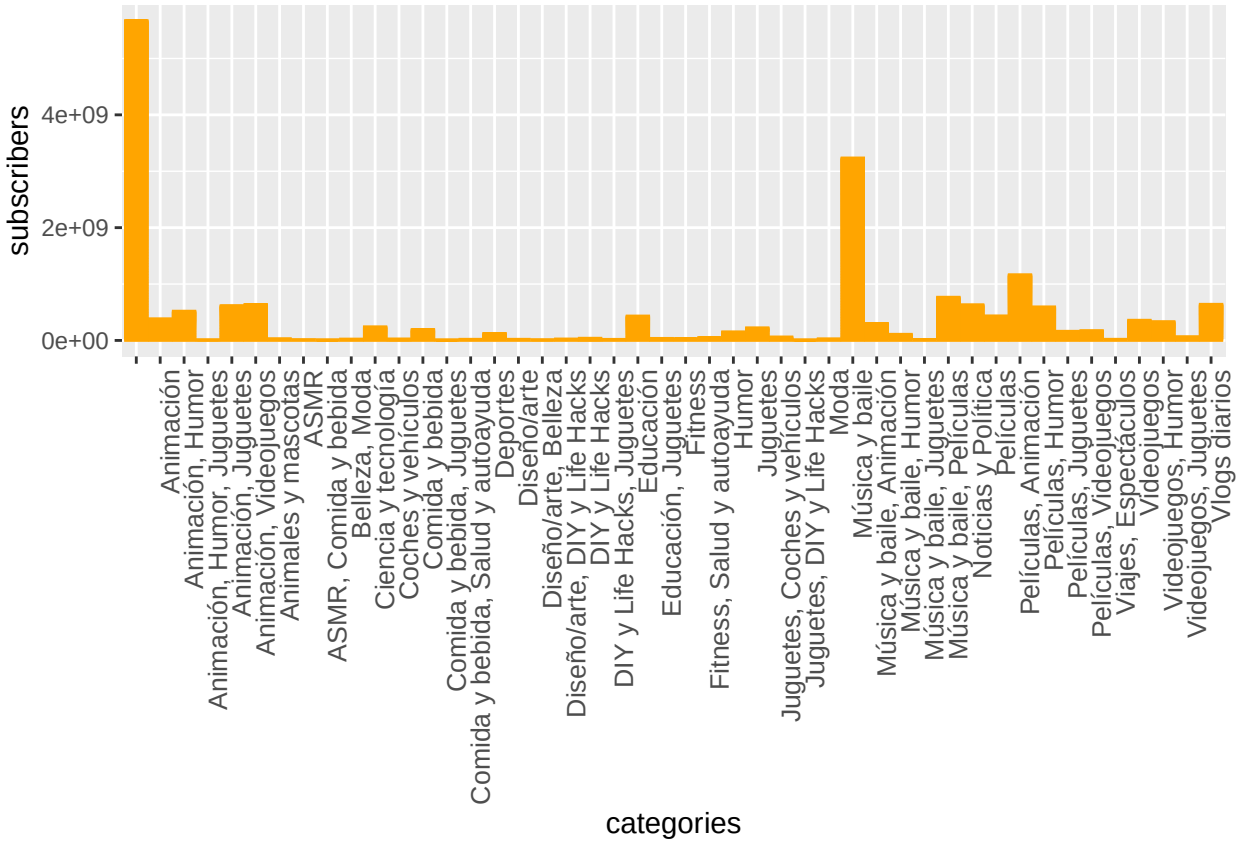
```
pe2
```


1.0e+08
7.5e+07
5.0e+07
2.5e+07
0.0e+00
visits
Animación
Animación, Humor
Animación, Humor, Juguetes
Animación, Juguetes
Animación, Videojuegos
Animales y mascotas
ASMR
ASMR, Comida y bebida
Belleza, Moda
Ciencia y tecnología
Coches y vehículos
Comida y bebida
Comida y bebida, Juguetes
Comida y bebida, Salud y autoayuda
Deportes
Diseño/arte
Diseño/arte, Belleza
Diseño/arte, DIY y Life Hacks
DIY y Life Hacks
DIY y Life Hacks, Juguetes
Educación
Educación, Juguetes
Fitness
Fitness, Salud y autoayuda
Humor
Juguetes
Juguetes, Coches y vehículos
Juguetes, DIY y Life Hacks
Moda
Música y baile
Música y baile, Animación
Música y baile, Humor
Música y baile, Juguetes
Música y baile, Películas
Noticias y Política
Películas
Películas, Animación
Películas, Humor
Películas, Juguetes
Películas, Videojuegos
Viajes, Espectáculos
Videojuegos
Videojuegos, Humor
Videojuegos, Juguetes
Vlogs diarios
categories

```
pe3
```


comments
30000
20000
10000
0
Animación
Animación, Humor
Animación, Humor, Juguetes
Animación, Juguetes
Animación, Videojuegos
Animales y mascotas
ASMR
ASMR, Comida y bebida
Belleza, Moda
Ciencia y tecnología
Coches y vehículos
Comida y bebida
Comida y bebida, Juguetes
Comida y bebida, Salud y autoayuda
Deportes
Diseño/arte
Diseño/arte, Belleza
Diseño/arte, DIY y Life Hacks
DIY y Life Hacks
DIY y Life Hacks, Juguetes
Educación
Educación, Juguetes
Fitness
Fitness, Salud y autoayuda
Humor
Juguetes
Juguetes, Coches y vehículos
Juguetes, DIY y Life Hacks
Moda
Música y baile
Música y baile, Animación
Música y baile, Humor
Música y baile, Juguetes
Música y baile, Películas
Noticias y Política
Películas
Películas, Animación
Películas, Humor
Películas, Juguetes
Películas, Videojuegos
Viajes, Espectáculos
Videojuegos
Videojuegos, Humor
Videojuegos, Juguetes
Vlogs diarios
categories

```
pe4
```



BRANDS AND COLLABORATIONS

The dataset does not have information about that so there is a need to create a proxy variables with performance metrics

```
ysa_numeric$brand_collaborations=ifelse(ysa_numeric$subscribers>18709023|
                                  ysa_numeric$visits>293544|
                                  ysa_numeric$likes>6292|
                                  ysa_numeric$comments>118,1,0)

brand_cor=cor(ysa_numeric)

kable(brand_cor, caption = "Correlations of perforance metrics")
```

Table 7: Correlations of perforance metrics

	rank	subscribers	visits	likes	comments	brand_collaborations
rank	1.0000000	-0.9653892	-0.0935175	-0.0266714	0.0223367	-0.4401339
subscribers	-0.9653892	1.0000000	0.0946686	0.0232043	-0.0280959	0.4577625

	rank	subscribers	visits	likes	comments	brand_collaborations
visits	-0.0935175	0.0946686	1.0000000	0.8173862	0.6546486	0.5473478
likes	-0.0266714	0.0232043	0.8173862	1.0000000	0.8154030	0.5598936
comments	0.0223367	-0.0280959	0.6546486	0.8154030	1.0000000	0.5291000
brand_collaborations	-0.4401339	0.4577625	0.5473478	0.5598936	0.5291000	1.0000000

- streamers with high number of performance metrics such as **likes** and **visits** are more likely to receive brand collaboration

BENCHMARKING

Top performing streamers in terms of likes

```
avg_likes=mean(ysa$likes)
avg_visits=mean(ysa$visits)
avg_comments=117
avg_subscribers=mean(ysa$subscribers)

top_streamers_likes=ysa %>%
  filter(likes > avg_likes)

top_streamers_likes=top_streamers_likes %>%
  select(c(username,likes))

top_streamers_likes=as.data.frame(top_streamers_likes)

top_streamers_likes=top_streamers_likes[order(-top_streamers_likes$likes),]

head(top_streamers_likes,10)
```

```
##                username    likes
## 1               MrBeast 13762.56
## 2             CoComelon 13762.56
## 4             PewDiePie 13762.56
## 5    LikeNastyaofficial 13762.56
## 6           VladandNiki 13762.56
## 8             BLACKPINK 13762.56
## 9                   BTS 13762.56
## 10           HYBELABELS 13762.56
## 11             ChuChuTV 13762.56
## 14 infobellshindirhymes 13762.56
```

Top 10 streamers in terms on subscribers

```
top_streamers_subscribers=ysa %>%
  filter(subscribers > avg_subscribers)

top_streamers_subscribers=top_streamers_subscribers %>%
```

```
  select(c(username,subscribers))

top_streamers_subscribers=as.data.frame(top_streamers_subscribers)

top_streamers_subscribers=top_streamers_subscribers[order(-top_streamers_subscribers$subscribers),]

head(top_streamers_subscribers,10)
```

```
##              username subscribers
## 1              tseries    28166020
## 2              MrBeast    28166020
## 3            CoComelon    28166020
## 4             SETIndia    28166020
## 5        KidsDianaShow    28166020
## 6            PewDiePie    28166020
## 7   LikeNastyaofficial    28166020
## 8          VladandNiki    28166020
## 9      zeemusiccompany    28166020
## 10                 WWE    28166020
```

Top 10 streamers in terms on visits

```
top_streamers_visits=ysa %>%
  filter(visits > avg_visits)

top_streamers_visits=top_streamers_visits %>%
  select(c(username,visits))

top_streamers_visits=as.data.frame(top_streamers_visits)

top_streamers_visits=top_streamers_visits[order(-top_streamers_visits$visits),]

head(top_streamers_visits,10)
```

```
##                username   visits
## 2             CoComelon 665338.9
## 3          KidsDianaShow 665338.9
## 4              PewDiePie 665338.9
## 5     LikeNastyaofficial 665338.9
## 6            VladandNiki 665338.9
## 13           dudeperfect 665338.9
## 14 infobellshindirhymes 665338.9
## 16           TaylorSwift 665338.9
## 17  BillionSurpriseToys 665338.9
## 18          ArianaGrande 665338.9
```

Top 10 streamers in terms of comments

```
top_streamers_comments=ysa %>%
  filter(comments > avg_comments)

top_streamers_comments=top_streamers_comments %>%
```

```
  select(c(username,comments))

top_streamers_comments=as.data.frame(top_streamers_comments)

top_streamers_comments=top_streamers_comments[order(-top_streamers_comments$comments),]

head(top_streamers_comments,10)
```

```
##        username comments
## 1       MrBeast 265.6684
## 2     PewDiePie 265.6684
## 4     BLACKPINK 265.6684
## 5           BTS 265.6684
## 6    HYBELABELS 265.6684
## 7   dudeperfect 265.6684
## 9   TaylorSwift 265.6684
## 10    EdSheeran 265.6684
## 11 ArianaGrande 265.6684
## 13 BillieEilish 265.6684
```

CONTENT RECOMMENDATIONS

A system for enhancing content recommendations to YouTube users based on streamers

```
streamer_metrics <- aggregate(cbind(visits, comments, likes, subscribers) ~ categories, ysa, mean)

normalized_metrics <- scale(streamer_metrics[, -1])

library(proxy)

similarity_matrix <- proxy::simil(normalized_metrics, method = "cosine")

s=streamer_metrics$categories

user_streamer <- s  # Streamers user has already interacted with

user_index <- which(streamer_metrics$categories == user_streamer)

similar_streamers <- order(similarity_matrix[user_index],decreasing = T)[-1]

recommended_streamers <- streamer_metrics$categories[similar_streamers[-1]]  # Exclude the user's own s

print(recommended_streamers)
```

```
##  [1] "Música y baile"                  "Videojuegos, Humor"
##  [3] "Noticias y Política"             "Moda"
##  [5] "Películas, Videojuegos"          "Música y baile, Juguetes"
##  [7] "Películas, Humor"                "Educación"
##  [9] "Películas"                       "Comida y bebida, Juguetes"
## [11] "Películas, Juguetes"             ""
## [13] "Deportes"                        "Música y baile, Humor"
## [15] "Música y baile, Películas"       "DIY y Life Hacks, Juguetes"
```

```
## [17] "Juguetes"                          "Educación, Juguetes"
## [19] "Diseño/arte, Belleza"              "Animación, Humor"
## [21] "ASMR, Comida y bebida"             "Comida y bebida"
## [23] "Juguetes, Coches y vehículos"      "Diseño/arte, DIY y Life Hacks"
## [25] "Animación"                         "Videojuegos, Juguetes"
## [27] "Videojuegos"                       "DIY y Life Hacks"
## [29] "Música y baile, Animación"         "Viajes, Espectáculos"
## [31] "Animación, Juguetes"               "Comida y bebida, Salud y autoayuda"
## [33] "Películas, Animación"              "Fitness"
## [35] "Animales y mascotas"               "Diseño/arte"
## [37] "Animación, Videojuegos"            "Juguetes, DIY y Life Hacks"
## [39] "Belleza, Moda"                     "Coches y vehículos"
## [41] "Fitness, Salud y autoayuda"        "ASMR"
## [43] "Vlogs diarios"                     "Ciencia y tecnología"
```

- The recommended youtube streamers belong to those categories.
- These categories help to classify streamers and provide a basis for recommending content to users with similar interests.

Key Findings

- Animacon is the most popular category with 306 streamers.
- Number of visits and likes have a string positive relationship.
- Estados Unidos is the country with the highest number of 293 streamers, followed by India with 241 streamers.
- Moda category has an exceptional performance metrics of more than 500 000 likes, 25 000 000 visits, more than 15 000 comments and 3 000 000 000 subscribers.
- Top 10 streamers have an average number of 13762 likes, 281 666 020 subscribers, 665338 visits and 205 comments

---