# YouTube Statistics 2023

Sheida Ebrahimi Siaghi

## Introduction

The dataset obtained from Kaggle providing information on the top 995 YouTube channels during 2023 offers a rich resource for exploring trends and patterns within the realm of YouTube creators. This dataset stands out due to its comprehensiveness, including details such as subscriber counts, views, video categories, country of origin, monthly earnings, and population data. Its breadth allows for multifaceted analyses, making it a valuable choice for investigating various aspects of YouTube creator dynamics.

In the following graphics, I'll delve into the top ten YouTube channels to uncover insights regarding their annual earnings, countries of origin, and most watched video categories. Additionally, I'll examine whether there's a correlation between the number of subscribers and the amount of money earned by these channels.

By exploring these questions, we can gain valuable insights into the landscape of YouTube content creation, identifying key trends and factors contributing to the success of top channels in 2023. Furthermore, this dataset's inclusion of population data alongside YouTube channel metrics offers a unique opportunity to contextualize the influence and reach of these top creators within their respective countries. Understanding the relationship between a channel's audience size, its revenue, and the population of its origin can provide deeper insights into the dynamics of content consumption and creator monetization strategies across different regions.

```
# Importing libraries

library(dplyr)
```

```
##
## 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
```

```
library(tidyr)
library(ggplot2)
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.3.3
```

```r
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.3.3
```

```r
library(scales)
```

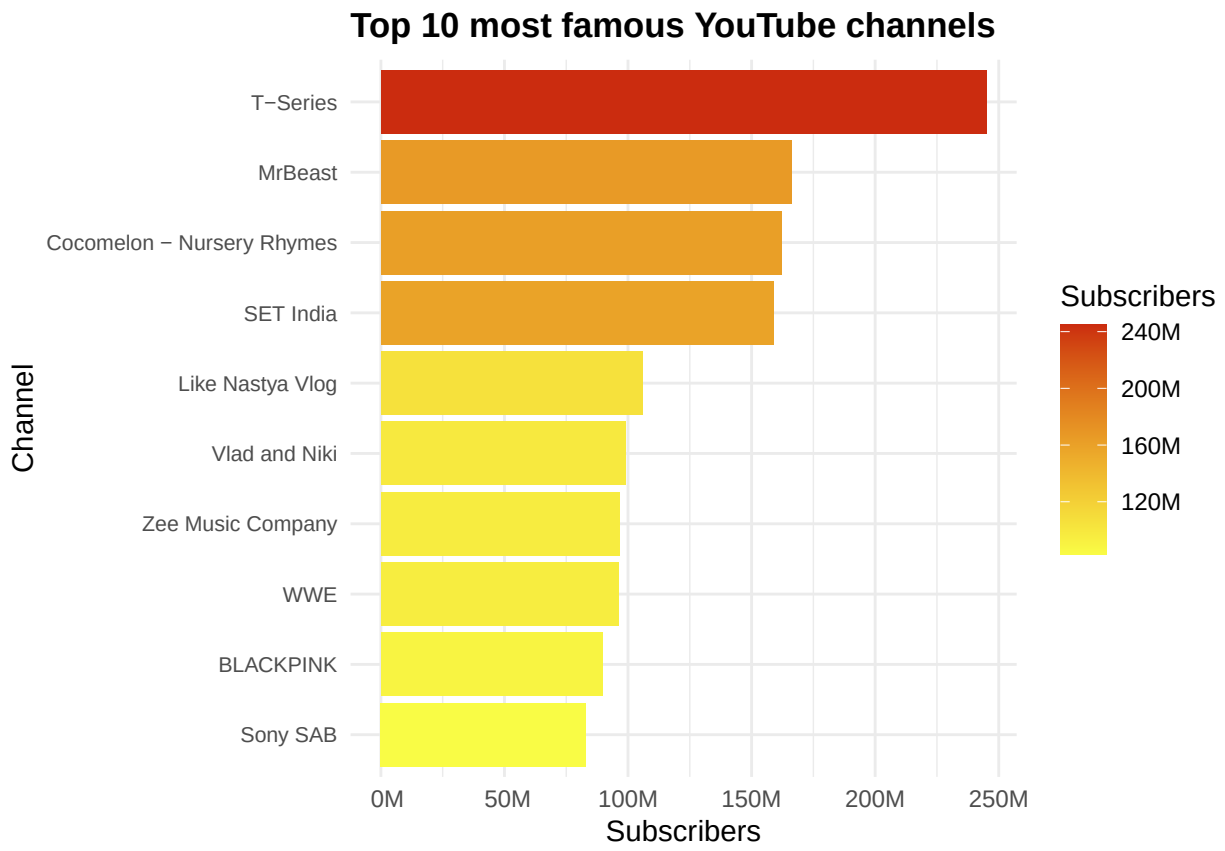
```r
# Loading the data
data <- read.csv("data.csv")
```

```r
# Cleaning the data of 'nan' values and making sure the countries match with our maps library
d <- data %>%
  mutate(across(where(is.numeric), ~ifelse(. == 'nan', NA, as.numeric(.))))
d <- d %>%mutate(across(where(is.character), ~ifelse(.== 'nan', NA, .)))
d <- na.omit(d)
d <- d %>%
  rename(Category = category)
d<- d %>%
  mutate(Country = case_when(
    Country == "United States" ~ "USA",
    Country == "United Kingdom" ~ "UK",
    TRUE ~ Country
  ))
```

Top 10 most famous YouTube channels can be viewed in the following chart:

```r
top_10_channels <- d %>%
  arrange(desc(subscribers)) %>%
  slice(1:10)

ggplot(top_10_channels, aes(x = reorder(Title, subscribers), y = subscribers, fill = subscribers)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 most famous YouTube channels", x = "Channel", y = "Subscribers") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8), plot.title = element_text(face = "bold")) +
  scale_y_continuous(labels = function(x) paste0(format(x / 1000000), "M")) +
  scale_fill_gradient(low = "#f9fc45", high = "#cb2c0f", name = "Subscribers",
                      labels = function(x) paste0(format(x / 1000000), "M"))
```



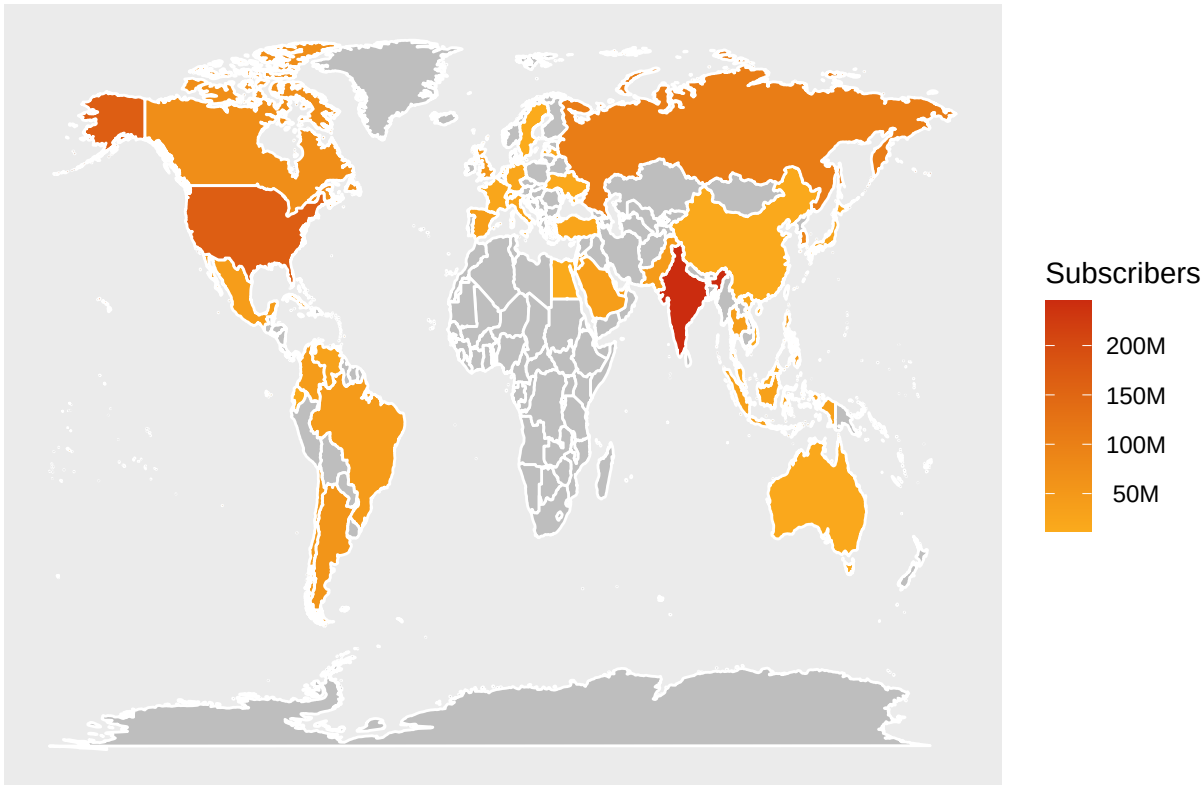

```
world_map <- map_data("world")
```

```
join <- left_join(world_map, d, by = c("region" = "Country"))
```

```
## Warning in left_join(world_map, d, by = c(region = "Country")): Detected an unexpected many-to-many r
## i Row 949 of `x` matches multiple rows in `y`.
## i Row 56 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

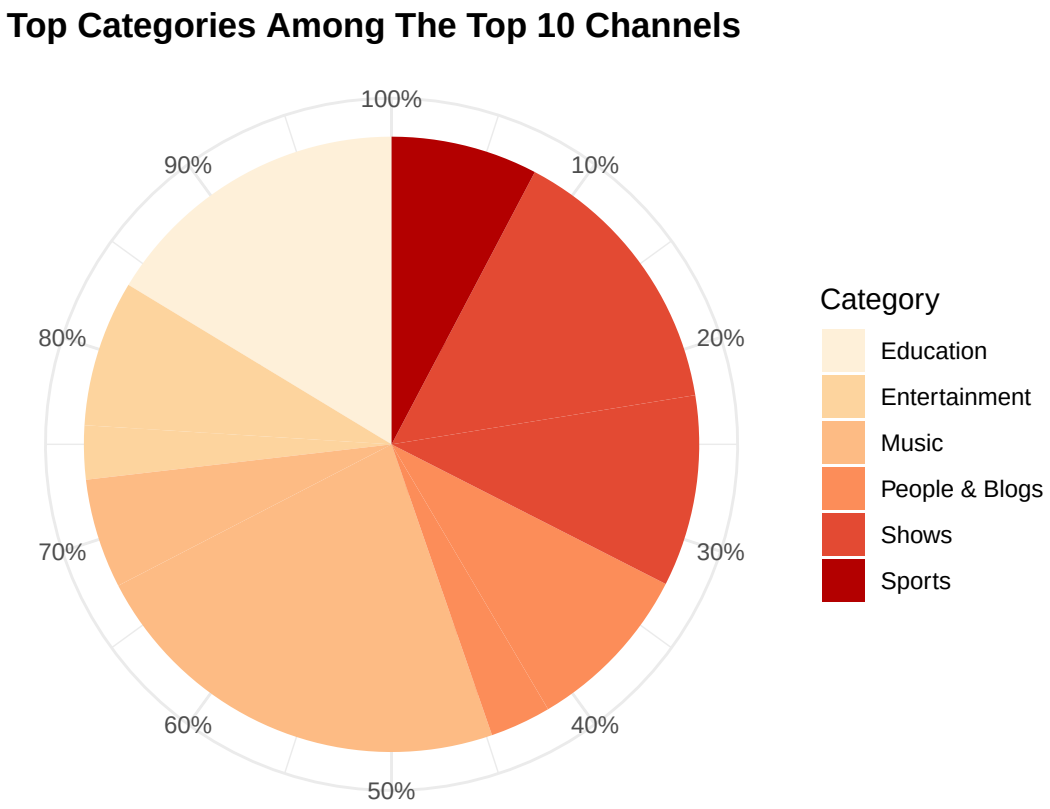
The country with the most subscribed to channels is India, followed by the United States and Russia.

```
ggplot(join, aes(x = long, y = lat, group = group, fill = subscribers)) +
  geom_polygon(color = "white") +
  scale_fill_gradient2(name = "Subscribers", labels = function(x) paste0(format(x / 1000000), "M"),
                       mid = "#fdb11d", high = "#cb2c0f", na.value = "grey") +
  labs(title = "Cumulative Subscribers Distribution Across Regions") +
  theme(panel.grid = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(face = "bold"))
```




Most viewed categories among famous channels were mainly Music and Sports.

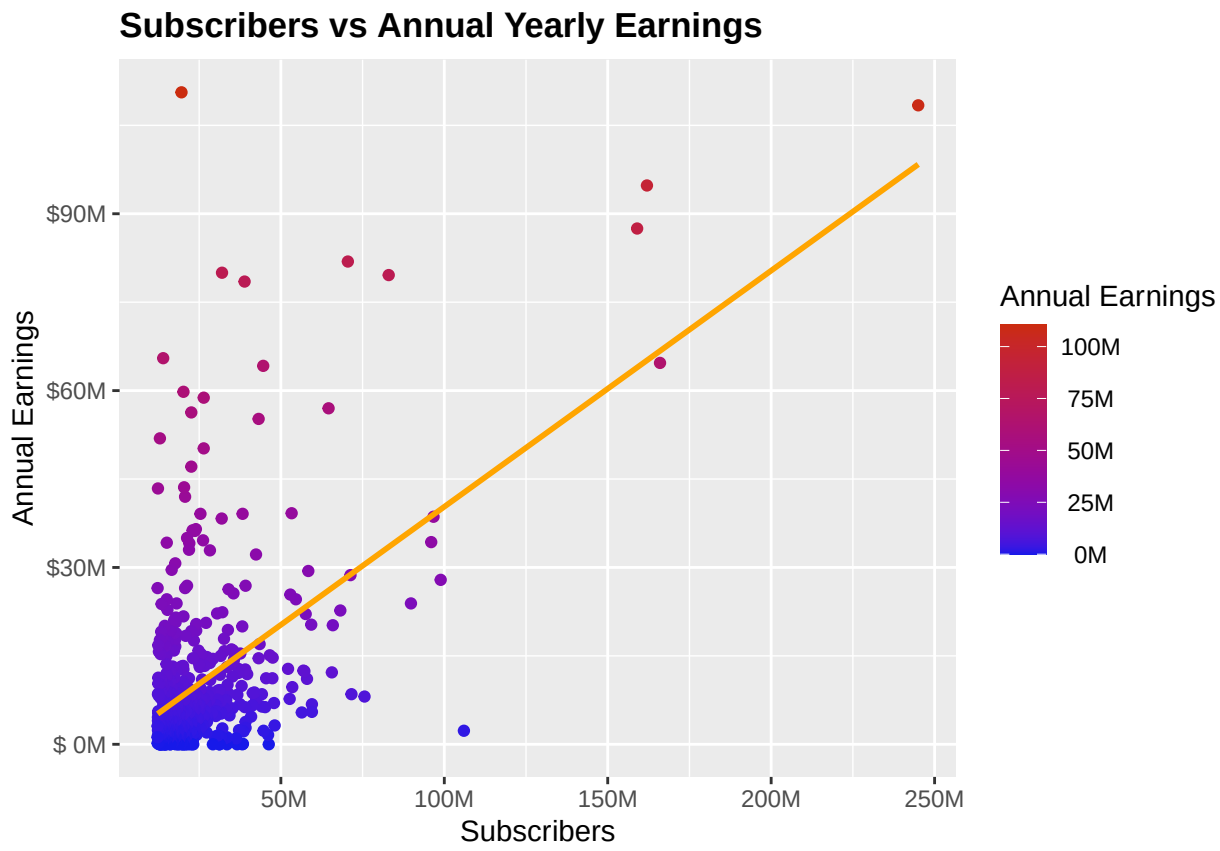
```
ggplot(top_10_channels, aes(x = "", y = video.views / sum(video.views), fill = Category)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  scale_fill_brewer(palette = "OrRd") +
  theme_minimal() +
  labs(title = "Top Categories Among The Top 10 Channels", colour = "") +
  theme(legend.position = "right", plot.title = element_text(face = "bold")) +
  scale_y_continuous(labels = scales::percent_format(scale = 100),breaks = seq(0.1, 1, by = 0.1)) +
  theme(legend.position = "right",
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks  = element_blank())
```




The number of subscribers directly influences a channel's earnings.

```
ggplot(d, aes(x = subscribers, y = highest_yearly_earnings, color = highest_yearly_earnings)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "#FFA500") +
  scale_x_continuous(labels = function(x) paste0(format(x / 1000000), "M"),
                     name = "Subscribers") +
  scale_y_continuous(labels = function(x) paste0("$", format(x / 1000000), "M"), name = "Annual Earnings
  scale_color_gradient(low = "#1719e9", high = "#cb2c0f", name = "Annual Earnings",
                       labels = function(x) paste0(format(x / 1000000), "M")) +
  labs(title = "Subscribers vs Annual Yearly Earnings") +
  theme(plot.title = element_text(face = "bold"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




## References

Source data - https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023