

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH (UEH)
TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ



ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

**ỨNG DỤNG BI VÀO PHÂN TÍCH VÀ ĐÁNH GIÁ
HIỆU SUẤT CỦA CÁC KÊNH YOUTUBE**

Học phần: Hệ Hỗ Trợ Quản Trị Thông Minh

Nhóm Sinh Viên:

- Nguyễn Đào Giang
- Võ Ngọc Mỹ Kim
- Nguyễn Thị Ngọc Nhi
- Văn Ngọc Như Quỳnh
- Nguyễn Ngọc Phương Trinh
- Nguyễn Nhật Thảo Vy

Chuyên Ngành: Khoa Học Dữ Liệu

Khóa: K47

Giảng Viên Hướng Dẫn: ThS. Phạm Thị Thanh Tâm

Thành phố Hồ Chí Minh, tháng 5 năm 2024

MỤC LỤC

MỤC LỤC.....	2
LỜI CẢM ƠN.....	5
CHƯƠNG 1: TỔNG QUAN	6
1.1. Lý do chọn đề tài.....	6
1.2. Mục đích, mục tiêu của đề tài	6
1.3. Đối tượng và phạm vi thực hiện	6
1.4. Phương pháp thực hiện	7
1.5. Bố cục của đề tài	7
1.6. Phân công công việc	7
CHƯƠNG 2: MÔ TẢ DOANH NGHIỆP.....	11
2.1. Giới thiệu doanh nghiệp.....	11
2.2. Thực trạng của doanh nghiệp.....	12
2.3. Bài toán của doanh nghiệp và mục tiêu cần giải quyết.....	14
2.3.1. Bài toán doanh nghiệp	14
2.3.2. Mục tiêu cần giải quyết.....	16
CHƯƠNG 3: QUÁ TRÌNH ETL.....	17
3.1. Giai đoạn 1: Trích xuất dữ liệu (Extract).....	17
3.1.1. Trích xuất dữ liệu	17
3.1.2. Mô tả bộ dữ liệu	17
3.1.3. Đánh giá chất lượng của dữ liệu	19
3.1.3.1. Độ tin cậy (Reliability)	19
3.1.3.2. Độ chính xác (Accuracy)	20
3.1.3.3. Độ thích hợp (Relevancy)	20
3.1.3.4. Độ kịp thời (Timeliness)	20
3.1.3.5. Sự đồng nhất (Consistency)	20
3.1.3.6. Sự phong phú (Richness)	20
3.1.3.7. Độ an toàn và Sự bảo mật (Security & Privacy).....	20
3.1.3.8. Khả năng tiếp cận (Accessibility)	20
3.1.4. Thống kê mô tả	21
3.1.4.1. Mô tả các biến định tính.....	21
3.1.4.2. Mô tả các biến định lượng	23
3.2. Giai đoạn 2: Chuyển đổi dữ liệu (Transform)	39
3.2.1. Thăm dò dữ liệu	39
3.2.2. Làm sạch dữ liệu	41
3.2.2.1. Đổi tên và loại bỏ cột	41
3.2.2.2. Loại bỏ ký tự lạ và các trùng lặp của cột Youtuber	42
3.2.2.3. Xử lý giá trị thiếu	44
3.2.2.4. Xử lý nhiễu	56
3.2.3. Tạo thêm thuộc tính	60

- Region.....	61
- Youtube Creator Awards	63
- Type of Youtuber.....	64
3.2.4. Thống kê mô tả sau tiền xử lý.....	68
3.2.4.1. Mô tả các biến định tính.....	68
3.2.4.2. Mô tả các biến định lượng	71
3.2.4.2.1. Độ phân tán.....	72
3.2.4.2.2. Xu thế trung tâm	73
3.2.5. Tạo các bảng dimension, fact.....	81
3.2.5.1. dimTime.....	81
3.2.5.2. dimYear.....	82
3.2.5.3. dimMonth.....	82
3.2.5.4. dimRegion.....	82
3.2.5.5. dimCountry	83
3.2.5.6. dimCategory.....	83
3.2.5.7. dimSubcategory	84
3.2.5.8. dimTypeofYoutuber.....	84
3.2.5.9. dimYoutuberCreatorAwards.....	85
3.2.5.10. factGYS.....	85
3.2.6. Chuyển đổi dữ liệu trong Power BI	86
3.2.6.1. Chính sửa tiêu đề cột.....	87
3.2.6.2. Thay đổi kiểu dữ liệu cột	88
3.3. Giai đoạn 3: Nạp dữ liệu vào kho dữ liệu (Load)	90
3.3.1. Giới thiệu các bảng Dim	90
3.3.2. Giới thiệu bảng Fact.....	94
CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU	99
4.1. Các Measure.....	99
4.2. Phân tích tổng quan về hiệu suất của các kênh Youtube (Overview).....	105
4.3. Phân tích hiệu suất của các kênh Youtube dựa theo năm (Analysis by Year)....	106
4.4. Phân tích hiệu suất của các kênh Youtube dựa theo tháng (Analysis by Month)	109
.....	
4.5. Phân tích hiệu suất của các kênh Youtube dựa theo vùng (Analysis by Region)	111
.....	
4.6. Phân tích hiệu suất của các kênh Youtube dựa theo quốc gia (Analysis by Country)	113
4.7. Phân tích hiệu suất của các kênh Youtube dựa theo chủ đề kênh (Analysis by Category).....	115
4.8. Phân tích hiệu suất của các kênh Youtube dựa theo chủ đề chi tiết của kênh (Analysis by Subcategory).....	116
4.9. Phân tích hiệu suất của các kênh Youtube dựa theo hình thức kênh youtube (Analysis by TypeofYoutuber)	118
4.10. Phân tích hiệu suất của các kênh Youtube dựa theo hình thức kênh youtube	

(Analysis by YoutubeCreatorAwards).....	120
CHƯƠNG 5: ĐỀ XUẤT GIẢI PHÁP	121
CHƯƠNG 6: KẾT LUẬN.....	128
TÀI LIỆU THAM KHẢO.....	129

LỜI CẢM ƠN

Kính gửi Thạc sĩ/Giảng viên Phạm Thị Thanh Tâm,

Nhóm 5 xin gửi lời cảm ơn chân thành nhất đến Cô vì đã tận tình hướng dẫn nhóm em trong quá trình học tập cũng như suốt quá trình thực hiện đồ án cuối kỳ môn Hệ hỗ trợ quản trị thông minh.

Sự tận tâm, kiến thức sâu rộng và những lời khuyên quý giá của Cô đã giúp nhóm rất nhiều trong việc hoàn thiện đồ án. Nhóm biết ơn sự kiên nhẫn của Cô khi giải đáp những thắc mắc của nhóm và luôn sẵn sàng dành thời gian để giúp đỡ nhóm.

Kính gửi các thành viên trong nhóm,

Em là Nguyễn Ngọc Phương Trinh - đại diện nhóm 5, cũng xin gửi lời cảm ơn chân thành đến các bạn trong nhóm đã cùng nhau nỗ lực hoàn thành đồ án này. Nhờ sự hợp tác, chia sẻ và hỗ trợ lẫn nhau của các bạn, nhóm đã học hỏi được rất nhiều điều và có những trải nghiệm quý báu.

Nhờ sự đóng góp của Cô và các bạn, đồ án của nhóm 5 đã được hoàn thành một cách tốt đẹp. Nhóm 5 xin hứa sẽ tiếp tục cố gắng học tập và rèn luyện để đạt được nhiều thành tích hơn nữa trong tương lai.

Trân trọng,

Nhóm 5

CHƯƠNG 1: TỔNG QUAN

1.1. Lý do chọn đề tài

YouTube là nền tảng chia sẻ video trực tuyến được ra mắt năm 2005 bởi Chad Hurley, Steve Chen và Jawed Karim. Sau khi được Google mua lại vào năm 2006, nền tảng này đã trở nên phổ biến trên toàn thế giới dưới sự nâng đỡ của ông lớn Google và thu hút một lượng rất lớn người dùng tìm hiểu và sử dụng nền tảng này để kiếm tiền, giải trí hoặc thu thập tin tức mới nhất trong ngày mà không phải đọc báo. Nổi bật nhất trong các đối thủ với YouTube hiện nay chính là TikTok. Trong khi YouTube tập trung vào video dài, nội dung đa dạng và tạo thu nhập từ quảng cáo và đối tác hóa, TikTok tập trung vào video ngắn, nội dung sáng tạo và tương tác xã hội. Với lợi thế hiện tại của mình là sự phổ biến và quy mô lớn, YouTube vẫn tiếp tục duy trì vị trí dẫn đầu và thu hút hàng tỷ lượt xem video mỗi ngày và thử thách lại đặt được vào tay các nhà sáng tạo nội dung. *Câu hỏi được đặt ra là làm sao để giữ chân người xem và tăng người dùng chính là bài toán cho các kênh đang hoạt động này?* Từ đó, việc phân tích và hiểu rõ các yếu tố ảnh hưởng đến sự quan tâm của người dùng cho các video và các kênh YouTube sẽ giúp các nhà sáng tạo nội dung duy trì và tăng cường sự quan tâm của công chúng đến nền tảng điện tử này.

1.2. Mục đích, mục tiêu của đề tài

Trong đề tài nghiên cứu, nhóm sẽ tập trung vào việc phân tích và đánh giá hiệu quả, hiệu suất của các kênh YouTube trên toàn thế giới. Nghiên cứu các yếu tố như số lượng người đăng ký, lượt xem, tần suất đăng video, thời lượng video, danh mục kênh và quốc gia để xác định các yếu tố quan trọng ảnh hưởng đến hiệu quả của kênh YouTube.Thêm vào đó, việc phân tích sự thay đổi về lượt xem, tương tác, số lượng người đăng ký, danh mục, sự ảnh hưởng của quốc gia đến hiệu quả và hiệu suất của các kênh YouTube, các xu hướng địa lý, phân bố và sự khác biệt giữa các quốc gia trong việc sử dụng và tiếp cận YouTube sẽ giúp đánh giá, đề xuất các chiến lược và phương pháp tối ưu hóa để nâng cao hiệu quả và hiệu suất của các kênh YouTube, tăng cường tương tác, tăng số lượng người đăng ký và tăng lượt xem.

1.3. Đối tượng và phạm vi thực hiện

Đối tượng nghiên cứu chính của nhóm là các kênh YouTube mà có số người đăng ký, lượt xem, tương tác cao trên nền tảng YouTube. Các kênh thuộc nhiều lĩnh vực và ngành nghề khác nhau, bao gồm cả cá nhân, nhóm và công ty trên nền tảng này. Đề

thực hiện phân tích chuyên sâu, nhóm sẽ sử dụng Python và Excel để thực hiện tiền xử lý dữ liệu. Sau đó, nhóm sẽ tiếp tục thực hiện các bước chuyển đổi và nạp dữ liệu vào Power BI để nghiên cứu và đánh giá bộ dữ liệu.

1.4. Phương pháp thực hiện

Nhóm thực hiện thu thập dữ liệu về các kênh youtube qua bộ dữ liệu *Global YouTube Statistics 2023* được tải về từ Kaggle. Xem xét các thông tin như tên kênh, số lượng người đăng ký, lượt xem, tương tác, danh mục, thời gian tạo kênh và các trường khác. Điều này giúp nhóm có cái nhìn tổng quan về dữ liệu và xác định các thuộc tính quan trọng.

Sau đó, tiếp tục thực hiện quá trình tiền xử lý các thông tin thu thập được ví dụ như xử lý nhiễu, loại bỏ các cột không cần thiết, giá trị trùng lặp và thêm một số thuộc tính để tiến hành phân tích sâu trước khi đưa dữ liệu vào Power BI.

Sau các bước tiền xử lý, nhóm tiến hành thêm các bước chuyển đổi dữ liệu như là về thuộc tính cột và tên cột. Tiếp theo tạo page phân tích tổng quan (overview) và thêm các page nhỏ để phân tích chuyên sâu bằng phương pháp khoan sâu. Trong quá trình thực hiện phân tích, nhóm cũng áp dụng kiến thức đã học để tạo thêm column và các measure tính toán khác để các dashboard thêm phần rõ ràng và trực quan.

1.5. Bố cục của đề tài

Chương 1: Giới thiệu tổng quan đề tài nghiên cứu

Chương 2: Giới thiệu tổng quan về doanh nghiệp và bài toán doanh nghiệp đối mặt để phân tích các hướng giải quyết

Chương 3: Thực hiện quá trình ETL bao gồm:

- E - Extract: Trích xuất dữ liệu *Global YouTube Statistics 2023* từ Kaggle.
- T - Transform: Chuyển đổi dữ liệu bao gồm làm sạch dữ liệu bằng cách xử lý các giá trị thiếu, loại bỏ dữ liệu trùng lặp hoặc không hợp lệ, chuẩn hóa dữ liệu và biến đổi các thuộc tính dữ liệu theo yêu cầu phân tích cụ thể khi đi vào chi tiết.
- L - Load: Sau khi dữ liệu đã được chuyển đổi, tiến hành nạp dữ liệu vào kho dữ liệu.

Chương 4: Thực hiện phân tích dữ liệu

Chương 5: Đề xuất các giải pháp cho các bài toán doanh nghiệp gấp phải

Chương 6: Đưa ra kết luận

1.6. Phân công công việc

MSSV	Thành viên	Phân công	Đánh giá
31211027636	Nguyễn Đào Giang	Tổng quan Thống kê mô tả (trước tiên xử lý) Giới thiệu các bảng dim Phân tích dựa trên TypeofYoutuber Phân tích dựa trên YoutubeCreatorAward Đề xuất, đánh giá giải pháp Kết luận	100%
31211027646	Võ Ngọc Mỹ Kim	Thống kê mô tả các biến Category (trước và sau tiền xử lý) Phân tích Country, Region Đề xuất, đánh giá giải pháp Thiết kế slide thuyết trình	100%
31211027660	Nguyễn Thị Ngọc Nhi	Trích xuất, mô tả, đánh giá bộ dữ liệu Thống kê mô tả (trước tiên xử lý) Tiền xử lý bộ dữ liệu Đề xuất, đánh giá giải pháp Tổng hợp báo cáo	100%
31211027669	Văn Ngọc Như Quỳnh	Mô tả doanh nghiệp Thống kê mô tả (trước tiên xử lý) Giới thiệu bảng Fact Phân tích dashboard Overview Hướng phát triển	100%

		Đề xuất, đánh giá giải pháp Tổng hợp dashboard	
31211026121	Nguyễn Ngọc Phương Trinh	Thống kê mô tả (trước tiền xử lý) Tạo bảng dim, fact Xây dựng mô hình dữ liệu Phân tích Category, Subcategory Đề xuất, đánh giá giải pháp Tổng hợp báo cáo	100%
31211025542	Nguyễn Nhật Thảo Vy	Thống kê mô tả biến numerics (trước và sau tiền xử lý) Phân tích Year, Month Đề xuất, đánh giá giải pháp	100%

Bảng 1 - Bảng phân công

CHƯƠNG 2: MÔ TẢ DOANH NGHIỆP

2.1. Giới thiệu doanh nghiệp



Hình 1 - Youtube logo

YouTube là một nền tảng truyền thông xã hội, nơi người dùng có thể xem, thích, chia sẻ, bình luận, chia sẻ và tải lên video của riêng họ. Được thành lập vào năm 2005, hiện nay YouTube là một trong những nền tảng phổ biến nhất trên mạng xã hội, với trung bình 694.000 giờ video được xem mỗi phút (thậm chí còn nhiều hơn cả Netflix).

Video đầu tiên được đăng tải trên YouTube vào ngày 23 tháng 4 năm 2005 bởi người đồng sáng lập YouTube - Jawed Karim, với video “Me at the zoo”. Đoạn phim 19 giây giờ đây đã trở thành huyền thoại, đã được xem hơn 317 triệu lượt.



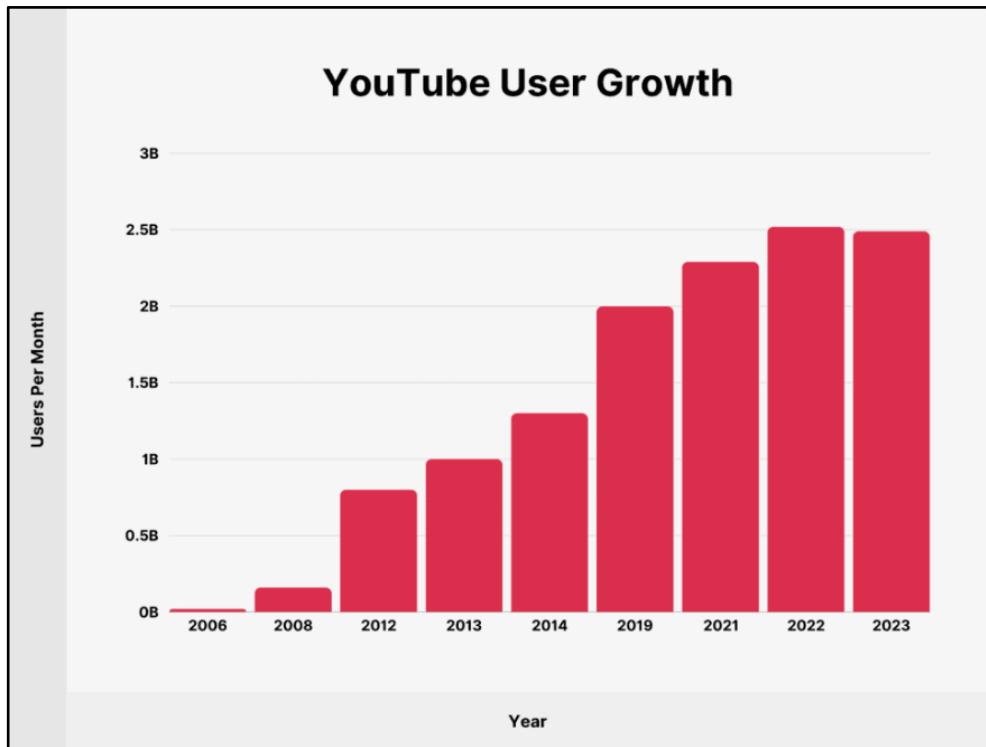
Hình 2 - Video “Me at the zoo”

Ngày nay, YouTube tận dụng những nhà sáng tạo nội dung, là cá nhân hoặc nhóm tạo nội dung đa dạng về giải trí, âm nhạc, thể thao, giáo dục và nhiều thể loại khác, giúp tạo nên một cộng đồng sáng tạo và đa dạng trên nền tảng này. Do vậy, YouTube luôn thu hút hàng tỷ người dùng để theo dõi những nhà sáng tạo nội dung yêu thích của họ hàng ngày để xem và tương tác.

Nhìn chung, YouTube vốn nổi tiếng với các nội dung video dài 10-30 phút, tuy nhiên, YouTube đã giới thiệu YouTube Shorts vào năm 2021. Giờ đây, các nhà sáng tạo nội dung trên YouTube có thể tải lên các video ngắn tối đa 60 giây (tương tự như Instagram Reels và TikToks).

2.2. Thực trạng của doanh nghiệp

- Tính đến năm 2023, YouTube có **2,49 tỷ người dùng** trên toàn thế giới.
- **47% người dùng internet** trên toàn cầu truy cập YouTube **hàng tháng** và **62% người dùng** truy cập YouTube **hàng ngày**.
- Người dùng xem hơn 1 tỷ giờ nội dung video YouTube mỗi ngày và dành trung bình 30 phút mỗi lần truy cập YouTube.
- YouTube **xếp thứ hai** trong danh sách các nền tảng truyền thông xã hội được sử dụng nhiều nhất, nền tảng được dùng nhiều nhất ngoài YouTube là Facebook của Mark Zuckerberg.
- Ấn Độ có lượng người dùng YouTube đông nhất, ước tính khoảng 462 triệu người. Tiếp theo là Hoa Kỳ, với 239 triệu người. Cho thấy đây là các thị trường tiềm năng nhất của YouTube.
- Các nhà sáng tạo nội dung nhận được trung bình từ 1,61 đến 29,3 đô la Mỹ cho mỗi 1000 lượt xem.



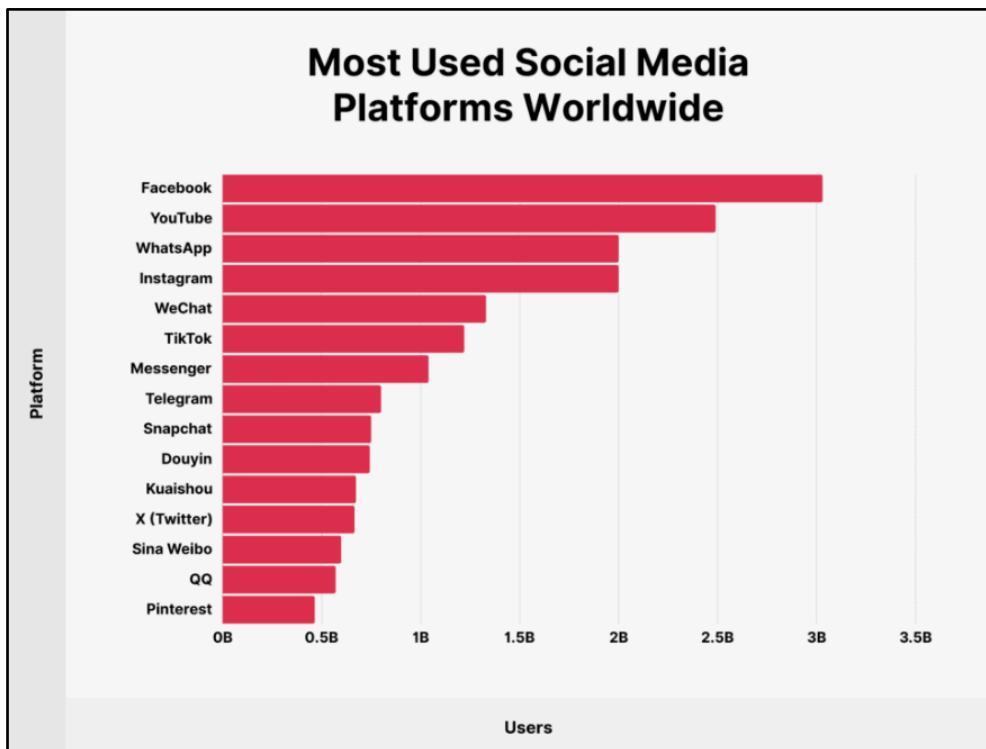
Hình 3 - Biểu đồ thể hiện số lượng người đăng ký các kênh trên YouTube (2006-2023)

Tham khảo: backlinko.com/youtube-users#usage-growth

(*) Biểu đồ không hiển thị số lượng người dùng YouTube đang hoạt động, chỉ hiển thị số lượng người đăng ký.

Từ biểu đồ trên, nhóm có thể đưa ra một số nhận xét sau:

- Nhìn chung, số lượng người đăng ký YouTube tăng đều đặn và ổn định trong suốt 19 năm qua.
- Năm 2006: Số lượng người đăng ký tăng sau khi Google mua lại YouTube.
- Năm 2012: Số lượng người đăng ký tăng đột biến sau khi YouTube phát hành ứng dụng của công ty trên iPhone, khiến YouTube là một trong ứng dụng được tải trước trên iPhone 5 và hệ điều hành iOS 6.
- Năm 2013, YouTube trở thành một trong các ứng dụng được dùng nhiều nhất.
- Năm 2021: Số lượng người đăng ký tăng sau khi YouTube giới thiệu tính năng video Shorts.



Hình 4 - Biểu đồ thể hiện số lượng người dùng trên các nền tảng xã hội (tính đến năm 2023)

Tham khảo: backlinko.com/youtube-users#most-used-social-media-platform

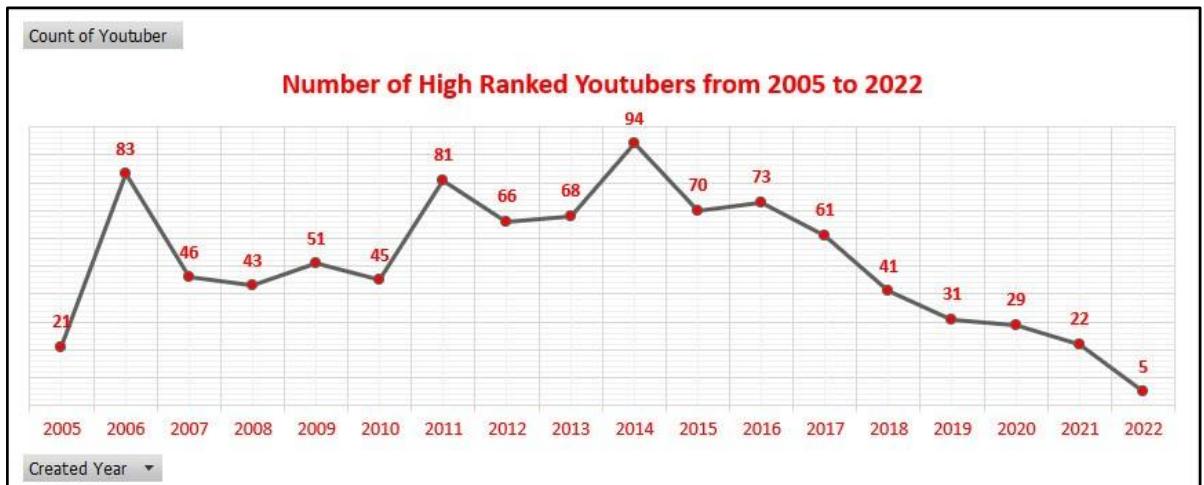
Từ biểu đồ trên, nhóm có thể đưa ra nhận xét như sau: Tính đến năm 2023, YouTube được xếp hạng thứ hai trong số tất cả các nền tảng truyền thông xã hội toàn cầu và chỉ đứng sau Facebook. Để dễ hình dung, trên toàn cầu chỉ có gần 4,95 tỷ người dùng mạng xã hội hoạt động. Do đó, YouTube là một trong hai nền tảng duy nhất có thể khẳng định nắm giữ ít nhất 50% thị phần người dùng trên toàn cầu.

*Tóm lại, hiện tại YouTube vẫn là một trong nền tảng truyền thông xã hội **đang phát triển mạnh mẽ nổi bật với số lượng người đăng ký ngày càng tăng**. YouTube có thể thu hút người dùng nhờ nội dung đa dạng, dễ dàng truy cập, tính tương tác và tiếp thị hiệu quả. Tuy nhiên, TikTok chỉ mất 6 năm để đạt đến 1,22 tỷ người dùng trong khi đó YouTube mất khoảng 9 năm để đạt được cột mốc tương tự - vì vậy, nếu với mức độ tăng trưởng đó, trong tương lai TikTok chắc chắn sẽ là một đối thủ tiềm năng đáng gờm và YouTube sẽ cần phải tiếp tục đổi mới để duy trì vị trí dẫn đầu của mình.*

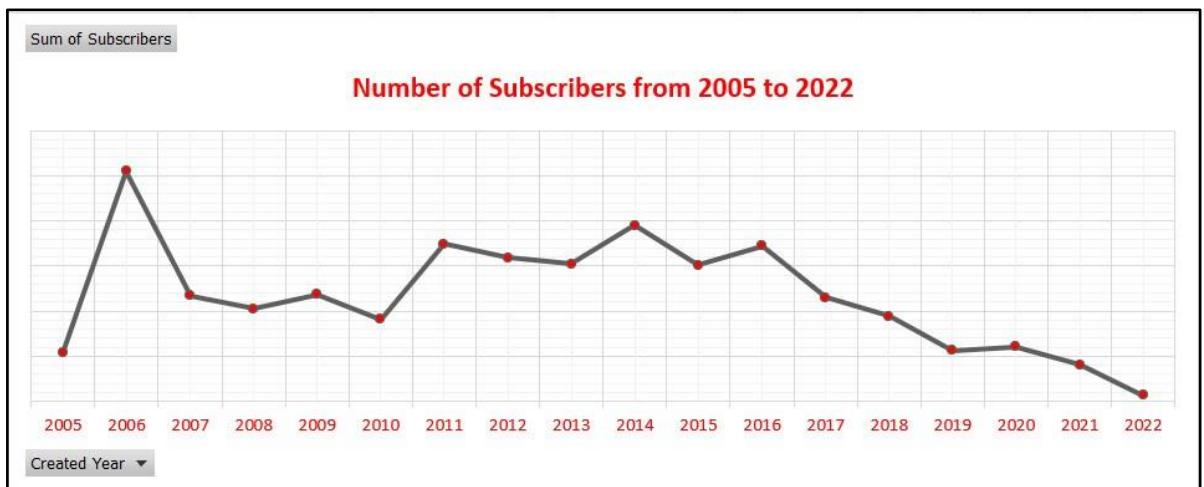
2.3. Bài toán của doanh nghiệp và mục tiêu cần giải quyết

2.3.1. Bài toán doanh nghiệp

Là ông vua trong lĩnh vực chia sẻ video trực tuyến, YouTube khẳng định vị thế dẫn đầu với sự phát triển vượt bậc so với các nền tảng mạng xã hội khác. Tuy nhiên, dựa trên bộ dữ liệu **Global YouTube Statistics 2023** mà nhóm đang nghiên cứu và phân tích thì có một tín hiệu đáng chú ý được ghi nhận là **số lượng YouTuber có thứ hạng cao trên bảng xếp hạng đang có xu hướng giảm dần qua từng năm** và **số lượng đăng ký của các kênh này cũng có xu hướng giảm** tương tự.



Hình 5 - Biểu đồ thể hiện số lượng kênh có thứ hạng cao trên YouTube (2005-2023)



Hình 6 - Biểu đồ thể hiện số lượng người đăng ký các kênh có thứ hạng cao trên YouTube (2005-2023)

Từ hai biểu đồ trên, nhóm có thể đưa ra một số nhận xét như sau:

- Số lượng YouTuber có thứ hạng cao tăng trưởng ổn định từ năm 2005 đến năm 2016, đạt đỉnh điểm vào năm 2016 với 94 người sáng tạo nội dung. Sau đó, số lượng này giảm dần đến năm 2022, chỉ còn 5 người sáng tạo nội dung.

- Số người đăng ký các kênh có thứ hạng cao thì cũng có xu hướng tương tự. Đat đỉnh điểm vào năm 2006 và từ năm 2007 đến năm 2016 thì biến động ổn định. Tuy nhiên, từ năm 2017 trở đi, số lượng này giảm đáng kể và giảm dần qua từng năm.

Điều này cho thấy sự cạnh tranh ngày càng gay gắt trên YouTube, khi mà việc bứt phá lên vị trí đỉnh cao trở nên khó khăn hơn bao giờ hết. Các yếu tố dẫn đến xu hướng này có thể bao gồm:

- Sự gia tăng của các nhà sáng tạo nội dung: Nền tảng thu hút ngày càng nhiều người tham gia sáng tạo nội dung, dẫn đến sự bão hòa và cạnh tranh cao hơn.
- Thay đổi thuật toán của YouTube: Các thay đổi trong thuật toán để xuất video có thể ảnh hưởng đến khả năng hiển thị của các kênh và tác động đến thứ hạng của họ.
- Xu hướng của người xem: Sở thích của người xem thay đổi liên tục, đòi hỏi các nhà sáng tạo nội dung phải thích nghi và đổi mới để duy trì lượng người theo dõi.

Ngoài ra còn có thể được giải thích bởi một số sự kiện đáng chú ý của YouTube như:

- Năm 2013, YouTube đã yêu cầu người dùng phải sử dụng Google+ để bình luận trên video, gây ra nhiều phản ứng tiêu cực từ cộng đồng YouTube. Dự án Google+ đã thất bại và đã ngừng hoạt động vào năm 2018, tạo ra sự bất ổn trong cộng đồng.
- Việc xử lý khiếu nại về vấn đề nội dung và bản quyền cũng gây ra nhiều sự tranh cãi. YouTube sẽ xóa toàn bộ video nếu chứa bất kỳ đoạn nhạc mẫu nhỏ nào được bảo vệ bản quyền, mặc dù sau này họ đã thay đổi cách tiếp cận bằng cách chuyển doanh thu quảng cáo từ video đó sang chủ sở hữu bản quyền.
- Năm 2019, YouTube đã đổi mới với nhiều khiếu nại về các bình luận được coi là những lời nói gây thù. Để giải quyết vấn đề này, YouTube đã hạn chế kiếm tiền và thậm chí xóa bỏ một số kênh, điều này đã gây ra nhiều chỉ trích.

Sự cạnh tranh gay gắt và chính sách kiểm duyệt khắt khe của YouTube đã và đang gây ra những tác động tiêu cực đến nền tảng này, đặc biệt là đối với hiệu suất hoạt động kênh các nhà sáng tạo nội dung và người dùng.

- Giảm số lượt xem: Do sự cạnh tranh cao, các video của họ có thể khó tiếp cận được với người xem hơn, dẫn đến giảm số lượt xem và doanh thu.
- Hạn chế bình luận: Các chính sách kiểm duyệt khắt khe của YouTube có thể dẫn đến việc hạn chế bình luận cho các video, khiến cho các nhà sáng tạo nội dung khó tương tác với người xem và nhận phản hồi từ họ. Có thể dẫn đến số lượng người đăng ký giảm đáng kể.

- Yêu cầu kiểm duyệt nội dung: Các nhà sáng tạo nội dung phải cẩn thận hơn với nội dung mà họ đăng tải, vì họ có thể bị xóa video hoặc thậm chí bị cấm khỏi nền tảng nếu vi phạm các quy định. *Các kênh có ít khả năng để trở thành những kênh thuộc top đầu.*

Do vậy, YouTube đang đổi mới với thách thức khi số lượng YouTuber có thứ hạng cao và số lượng đăng ký các kênh đó giảm dần qua từng năm.

2.3.2. Mục tiêu cần giải quyết

Để giải quyết những thách thức này, YouTube cần đặt ra những mục tiêu sau:

- Tăng số lượng YouTuber có thứ hạng cao
- Tăng số lượng đăng ký kênh
- Tăng cường khả năng cạnh tranh giữa YouTube so với các nền tảng khác

CHƯƠNG 3: QUÁ TRÌNH ETL

3.1. Giai đoạn 1: Trích xuất dữ liệu (Extract)

- Link nguồn: [Global Youtube Statistics 2023](#)

3.1.1. Trích xuất dữ liệu

Global Youtube Statistics 2023 được đăng tải bởi Nidula Elgiriyewithna trên Kaggle vào tháng 07 năm 2023. Mục đích của bộ dữ liệu này là phân tích và thu thập thông tin chi tiết về những người sáng tạo nội dung nổi tiếng trên Youtube, nền tảng video trực tuyến phổ biến nhất thế giới.

Bộ dữ liệu cung cấp các số liệu thống kê chi tiết về các kênh Youtube có lượng đăng ký nhiều nhất từ năm 2005 đến 2023. Đây là một tập hợp về các kênh hàng đầu trên Youtube, tạo ra những cơ hội để nghiên cứu và hiểu rõ hơn về những người dẫn đầu trên nền tảng này. Bộ dữ liệu gồm 28 thuộc tính và 995 dòng, với các thuộc tính nổi bật như số lượng người đăng ký, lượt xem video, tần suất đăng video, thu nhập của

các nhà sáng tạo nội dung hàng đầu.

Với việc phân tích các thuộc tính này, bộ dữ liệu này sẽ trở thành một nguồn thông tin vô giá cho các nhà sáng tạo nội dung tiềm năng, những người đam mê dữ liệu và bất kỳ ai quan tâm đến lĩnh vực nội dung trực tuyến đang không ngừng phát triển. Bộ dữ liệu này không chỉ cung cấp thông tin mà còn là một công cụ vô cùng quý giá để hiểu rõ hơn về sự phát triển và xu hướng của nội dung trực tuyến.

3.1.2. Mô tả bộ dữ liệu

Tên thuộc tính	Mô tả	Kiểu dữ liệu
rank	Thứ hạng của các kênh trên Youtube	int64
Youtuber	Tên của Youtuber	object
subscribers	Số lượng người đăng ký kênh	int64
video views	Tổng số lượt xem trên tất cả các video của kênh	float64 4
category	Danh mục của kênh	object
Title	Tên của kênh Youtube	object
uploads	Tổng số lượng video được đăng tải trên kênh	int64
Country	Quốc gia mà kênh được tạo	object
Abbreviation	Tên viết tắt của quốc gia	object

channel_type	Chủ đề của kênh	object
video_views_rank	Xếp hạng kênh dựa vào tổng lượt xem	float6 4
country_rank	Xếp hạng kênh dựa vào số lượng người đăng ký trong nước	float6 4
channel_type_rank	Xếp hạng kênh dựa vào chủ đề của kênh	float6 4
video_views_for_the_last_30_days	Tổng lượt xem video trong 30 ngày qua	float6 4
lowest_monthly_earnings	Ước tính thu nhập hàng tháng thấp nhất từ kênh	float6 4
highest_monthly_earnings	Ước tính thu nhập hàng tháng cao nhất từ kênh	float6 4
lowest_yearly_earnings	Ước tính thu nhập hàng năm thấp nhất từ kênh	float6 4
highest_yearly_earnings	Ước tính thu nhập hàng năm cao nhất từ kênh	float6 4
subscribers_for_last_30_days	Số lượng người mới đăng ký trong 30 ngày qua	float6 4
created_year	Năm mà kênh được tạo	float6 4
created_month	Tháng mà kênh được tạo	object
created_date	Ngày mà kênh được tạo	float6 4

Gross tertiary education enrollment (%)	Tỷ lệ dân số theo học đại học trong nước	float6 4
Population	Tổng dân số của quốc gia	float6 4
Unemployment rate	Tỷ lệ thất nghiệp của quốc gia	object
Urban_population	Tỷ lệ dân số sống ở khu vực đô thị	float6 4
Latitude	Vĩ độ của quốc gia	object
Longitude	Kinh độ của quốc gia	object

Bảng 2 - Mô tả bộ dữ liệu gốc

3.1.3. Đánh giá chất lượng của dữ liệu

3.1.3.1. Độ tin cậy (Reliability)

Bộ dữ liệu **Global Youtube Statistics 2023** được thu thập từ Kaggle, một nền tảng uy tín cho các nhà khoa học dữ liệu và cộng đồng học máy. Nidula Elgiriyewithna, người đăng tải bộ dữ liệu, có uy tín trong lĩnh vực phân tích dữ liệu Youtube.

3.1.3.2. Độ chính xác (Accuracy)

Bộ dữ liệu được thu thập từ các nguồn chính thức của Youtube, bao gồm API Youtube và trang web Youtube. Tuy nhiên, vẫn có khả năng xảy ra lỗi do quá trình thu thập dữ liệu tự động.

3.1.3.3. Độ thích hợp (Relevancy)

Bộ dữ liệu bao gồm nhiều thông tin chi tiết về kênh Youtube và video như ngày tạo kênh, số lượt xem và người đăng ký, phù hợp cho mục đích phân tích và nghiên cứu về Youtube.

3.1.3.4. Độ kịp thời (Timeliness)

Bộ dữ liệu được cập nhật vào tháng 07 năm 2023, tuy nhiên một số thông tin có

thể đã thay đổi.

3.1.3.5. Sự đồng nhất (Consistency)

Bộ dữ liệu được định dạng tốt và có cấu trúc rõ ràng gồm các dòng và cột, và đều chỉ được thu thập duy nhất trên nền tảng Youtube.

3.1.3.6. Sự phong phú (Richness)

Bộ dữ liệu bao gồm nhiều thông tin chi tiết, bao gồm thông tin về các kênh Youtube, video, và người sáng tạo nội dung trên toàn thế giới và có đa dạng lĩnh vực, thể loại.

3.1.3.7. Độ an toàn và Sự bảo mật (Security & Privacy)

Bộ dữ liệu không bao gồm thông tin cá nhân của người dùng Youtube.

Ngoài tác giả Nidula Elgiriyewithna thì không ai có quyền chỉnh sửa các thông tin của bộ dữ liệu. Việc hạn chế quyền truy cập và chỉnh sửa dữ liệu giúp đảm bảo tính toàn vẹn và chính xác của bộ dữ liệu.

3.1.3.8. Khả năng tiếp cận (Accessibility)

Bộ dữ liệu có thể truy cập miễn phí trên Kaggle.

Tóm lại, bộ dữ liệu **Global Youtube Statistics 2023** là một bộ dữ liệu chất lượng cao, phù hợp cho mục đích phân tích và nghiên cứu về Youtube. Tuy nhiên, nhóm cũng có lưu ý rằng bộ dữ liệu này có thể không hoàn toàn chính xác và cập nhật.

Tiêu chí	Đánh giá
Độ tin cậy	Cao
Độ chính xác	Trung bình
Độ thích hợp	Cao
Độ kịp thời	Trung bình
Sự đồng nhất	Cao
Sự phong phú	Cao
Độ an toàn và Sự bảo mật	Cao

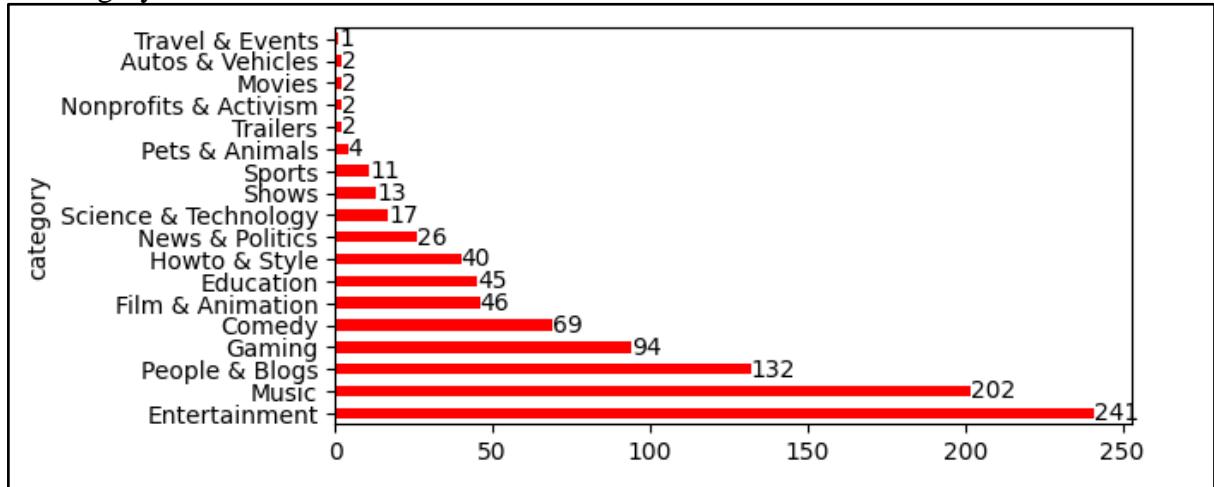
mật	
Khả năng tiếp cận	Cao

Bảng 3 - Đánh giá chất lượng dữ liệu

3.1.4. Thống kê mô tả

3.1.4.1. Mô tả các biến định tính

3.1.4.1.1. Category

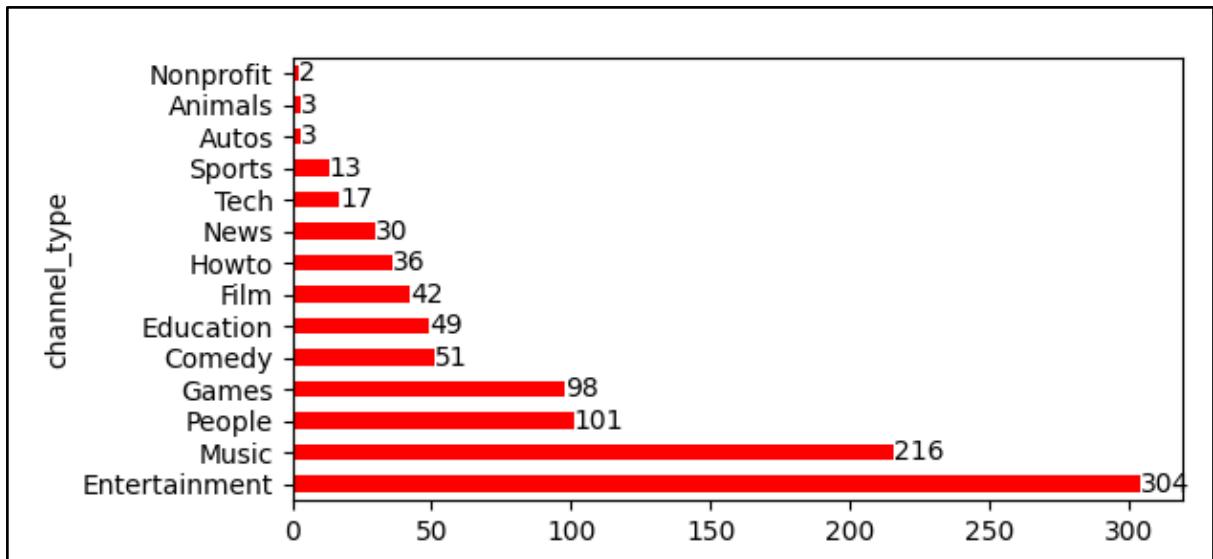


Hình 7 - Biểu đồ thể hiện tần số theo biến category (trước tiền xử lý)

Nhân xét:

- Các kênh youtube có đa dạng thể loại đến từ nhiều chủ đề khác nhau như du lịch, giáo dục, giải trí đến động vật, thể thao.
- Bộ dữ liệu ghi nhận được tổng cộng có 18 thể loại khác nhau.
- Những thể loại mà được các nhà sáng tạo nội dung chọn lựa nhiều nhất lần lượt là Entertainment (241), Music (202) và People & Blogs (132). Điều này cũng cho thấy xu hướng hiện nay là làm các nội dung về giải trí, âm nhạc, con người và đời sống hơn là các hoạt động phi lợi nhuận (Nonprofits & Activism) (2) và tự động hóa (Autos & Vehicles) (2).

3.1.4.1.2. Channel Type

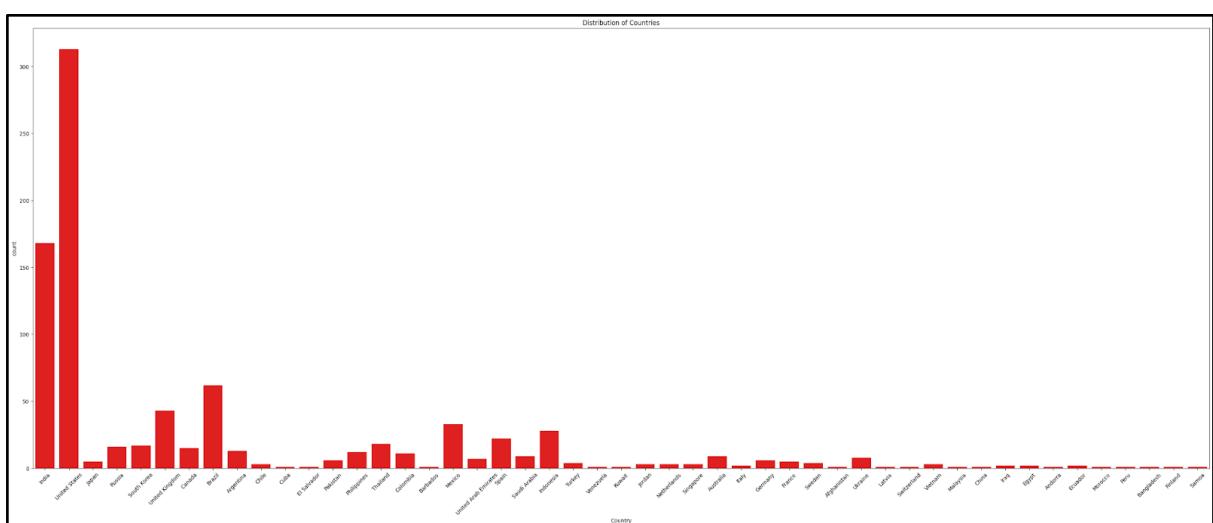


Hình 8 - Biểu đồ thể hiện tần số theo biến channel - type (trước tiền xử lý)

Nhận xét:

- Cũng tương như thể loại, các kênh youtube cũng có đa dạng chủ đề khác nhau như từ công nghệ, giáo dục, giải trí đến động vật, thể thao.
- Bộ dữ liệu ghi nhận được tổng cộng có 14 chủ đề khác nhau.
- Những chủ đề mà được các nhà sáng tạo nội dung chọn lựa nhiều nhất lần lượt là Entertainment (304), Music (216) và People (101). Điều này cũng cho thấy xu hướng hiện nay là làm các nội dung về giải trí, âm nhạc, con người và đời sống hơn là các chủ đề về hoạt động phi lợi nhuận (Nonprofits) (2) hay động vật (Animals) (3).

3.1.4.1.3. Country



Hình 9 - Biểu đồ thể hiện tần số theo biến country (trước tiền xử lý)

Nhận xét:

- Bộ dữ liệu ghi nhận được các kênh youtube được tạo từ 48 quốc gia khác nhau. Điều này cho thấy sự đa dạng của nền tảng Youtube, cũng như sự phổ biến của nền tảng này.
- Số lượng Youtuber đến từ các quốc gia này cũng có sự chênh lệch khá lớn. Như số lượng youtuber từ Mỹ (United States) và Ấn Độ (India) chiếm hơn một nửa bộ dữ liệu.

3.1.4.2. Mô tả các biến định lượng

3.1.4.2.1. Rank

Các đại lượng về xu thế trung tâm của biến rank

Mean: 498.0

Mode: [None]

Median: 498.0

Các đại lượng về độ phân tán biến rank

Khoảng biến thiên (Range): 994

Phương sai (Variance): 82585.0

Độ lệch chuẩn (Standard deviation): 287.376

Các đại lượng về hình dáng phân phối biến rank

Độ lệch: 0.0

Độ nhọn: -1.2

Nhận xét:

- Các đại lượng về xu thế trung tâm và về độ phân tán không có ý nghĩa phân tích đối với biến này.

3.1.4.2.2. Subscribers

Các đại lượng về xu thế trung tâm của biến subscribers

Mean: 22982412.06

Mode: [12500000]

Median: 17700000.0

Các đại lượng về độ phân tán biến subscribers

Khoảng biến thiên (Range): 232700000

Phương sai (Variance): 307164368421584.75

Độ lệch chuẩn (Standard deviation): 17526105.341

Các đại lượng về hình dáng phân phối biến subscribers

Độ lệch: 5.51

Độ nhọn: 45.51

Nhận xét:

- Biến subscribers có giá trị trung bình là 22982412.06, trung vị là 17700000.0 và giá trị thường xuất hiện nhất là 12500000.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 5.51, giá trị độ lệch dương cho thấy phân phối subscribers bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng subscribers thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng subscribers cao hơn nhiều so với giá trị trung bình; Độ nhọn là 45.51, giá trị độ nhọn cao cho thấy phân phối subscribers có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.3. Video views

Các đại lượng về xu thế trung tâm của biến video views

Mean: 11039537052.038

Mode: [0.0]

Median: 7760819588.0

Các đại lượng về độ phân tán biến video views

Khoảng biến thiên (Range): 228000000000.0

Phương sai (Variance): 1.991159290271055e+20

Độ lệch chuẩn (Standard deviation): 14110844376.83

Các đại lượng về hình dáng phân phối biến video views

Độ lệch: 7.2

Độ nhọn: 82.54

Nhận xét:

- Biến video_views có giá trị trung bình là 11039537052.038, trung vị là 7760819588.0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 7.2, giá trị độ lệch dương cho thấy phân phối video_views bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng video_views thấp hơn giá trị trung bình, và

một số ít đối tượng có số lượng video_views cao hơn nhiều so với giá trị trung bình; Độ nhọn là 82.54, giá trị độ nhọn cao cho thấy phân phối video_views có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.4. Uploads

Các đại lượng về xu thế trung tâm của biến uploads
Mean: 9187.126
Mode: [0]
Median: 729.0

Các đại lượng về độ phân tán biến uploads
Khoảng biến thiên (Range): 301308
Phương sai (Variance): 1166314860.758
Độ lệch chuẩn (Standard deviation): 34151.352

Các đại lượng về hình dáng phân phối biến uploads
Độ lệch: 5.66
Độ nhọn: 35.44

Nhận xét:

- Biến uploads có giá trị trung bình là 9187.126, trung vị là 729.0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 5.66, giá trị độ lệch dương cho thấy phân phối uploads bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng uploads thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng uploads cao hơn nhiều so với giá trị trung bình; Độ nhọn là 35.44, giá trị độ nhọn cao cho thấy phân phối uploads có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.5. Video views rank

Các đại lượng về xu thế trung tâm của biến video_views_rank
Mean: 554248.904
Mode: [4057944.0]
Median: nan

Các đại lượng về độ phân tán biến video_views_rank

Khoảng biến thiên (Range): nan

Phương sai (Variance): 1857175351568.848

Độ lệch chuẩn (Standard deviation): 1362782.21

Các đại lượng về hình dáng phân phối biến video_views_rank

Độ lệch: nan

Độ nhọn: nan

Nhận xét:

- Biến video_views_rank có giá trị trung bình là 554248.904 và giá trị thường xuất hiện nhất là 4057944.0.
- Dữ liệu có mức độ phân tán rộng, thể hiện giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến video_views_rank.

3.1.4.2.6. Country rank

Các đại lượng về xu thế trung tâm của biến country_rank

Mean: 386.053

Mode: [1.0]

Median: nan

Các đại lượng về độ phân tán biến country_rank

Khoảng biến thiên (Range): nan

Phương sai (Variance): 1518427.114

Độ lệch chuẩn (Standard deviation): 1232.245

Các đại lượng về hình dáng phân phối biến country_rank

Độ lệch: nan

Độ nhọn: nan

Nhận xét:

- Biến country_rank có giá trị trung bình là 386.053 và giá trị thường xuất hiện nhất là 1.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến country_rank.

3.1.4.2.7. Channel type rank

Các đại lượng về xu thế trung tâm của biến channel_type_rank
Mean: 745.719
Mode: [1.0]
Median: nan

Các đại lượng về độ phân tán biến channel_type_rank
Khoảng biến thiên (Range): nan
Phương sai (Variance): 3780639.099
Độ lệch chuẩn (Standard deviation): 1944.387

Các đại lượng về hình dáng phân phối biến channel_type_rank
Độ lệch: nan
Độ nhọn: nan

Nhân xét:

- Biến channel_type_rank có giá trị trung bình là 745.719 và giá trị thường xuất hiện nhất là 1.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến channel_type_rank.

3.1.4.2.8. Video views for the last 30 days

Các đại lượng về xu thế trung tâm của biến video_views_for_the_last_30_days
Mean: 175610284.794
Mode: [1.0]
Median: nan

Các đại lượng về độ phân tán biến video_views_for_the_last_30_days
Khoảng biến thiên (Range): nan
Phương sai (Variance): 1.7337077729358685e+17
Độ lệch chuẩn (Standard deviation): 416378166.207

Các đại lượng về hình dáng phân phối biến video_views_for_the_last_30_days
Độ lệch: nan
Độ nhọn: nan

Nhân xét:

- Biến video_views_for_the_last_30_days có giá trị trung bình là 175.610.284,794 và giá trị thường xuất hiện nhất là 1.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến video_views_for_the_last_30_days.

3.1.4.2.9. Subscribers for the last 30 days

Các đại lượng về xu thế trung tâm của biến subscribers_for_last_30_days

Mean: 349079.132

Mode: [100000.0]

Median: nan

Các đại lượng về độ phân tán biến subscribers_for_last_30_days

Khoảng biến thiên (Range): nan

Phương sai (Variance): 377432607917.36

Độ lệch chuẩn (Standard deviation): 614355.441

Các đại lượng về hình dáng phân phối biến subscribers_for_last_30_days

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến subscribers_for_last_30_days có giá trị trung bình là 349.079,132 và giá trị thường xuất hiện nhất là 100.000.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến subscribers_for_last_30_days.

3.1.4.2.10. Lowest monthly earnings

Các đại lượng về xu thế trung tâm của biến lowest_monthly_earnings

Mean: 36886.148

Mode: [0.0]

Median: 13300.0

Các đại lượng về độ phân tán biến lowest_monthly_earnings

Khoảng biến thiên (Range): 850900.0

Phương sai (Variance): 5163676228.168

Độ lệch chuẩn (Standard deviation): 71858.724

Các đại lượng về hình dáng phân phối biến lowest_monthly_earnings

Độ lệch: 4.79

Độ nhọn: 31.86

Nhân xét:

- Biến lowest_monthly_earnings có giá trị trung bình là 36.886,148, trung vị là 13.300,0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 4.79, giá trị độ lệch dương cho thấy phân phối lowest_monthly_earnings bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng lowest_monthly_earnings thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng lowest_monthly_earnings cao hơn nhiều so với giá trị trung bình; Độ nhọn là 31.85, giá trị độ nhọn cao cho thấy phân phối lowest_monthly_earnings có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.11. Highest monthly earnings

Các đại lượng về xu thế trung tâm của biến highest_monthly_earnings

Mean: 589807.848

Mode: [0.0]

Median: 212700.0

Các đại lượng về độ phân tán biến highest_monthly_earnings

Khoảng biến thiên (Range): 13600000.0

Phương sai (Variance): 1319333598065.873

Độ lệch chuẩn (Standard deviation): 1148622.478

Các đại lượng về hình dáng phân phối biến highest_monthly_earnings

Độ lệch: 4.79

Độ nhọn: 31.82

Nhân xét:

- Biến highest_monthly_earnings có giá trị trung bình là 589.807,848, trung vị là 212.700,0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 4.79, giá trị độ lệch dương cho thấy phân phối highest_monthly_earnings bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng highest_monthly_earnings thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng highest_monthly_earnings cao hơn nhiều so với giá trị trung bình; Độ nhọn là 31.82, giá trị độ nhọn cao cho thấy phân phối highest_monthly_earnings có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.12. Lowest yearly earning

Các đại lượng về xu thế trung tâm của biến lowest_yearly_earnings

Mean: 442257.393

Mode: [0.0]

Median: 159500.0

Các đại lượng về độ phân tán biến lowest_yearly_earnings

Khoảng biến thiên (Range): 10200000.0

Phương sai (Variance): 741693177690.002

Độ lệch chuẩn (Standard deviation): 861216.104

Các đại lượng về hình dáng phân phối biến lowest_yearly_earnings

Độ lệch: 4.79

Độ nhọn: 31.94

Nhân xét:

- Biến lowest_yearly_earnings có giá trị trung bình là 442257.393, trung vị là 159500.0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.

- Độ lệch là 4.79, giá trị độ lệch dương cho thấy phân phối lowest_yearly_earnings bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng lowest_yearly_earnings thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng lowest_yearly_earnings cao hơn nhiều so với giá trị trung bình; Độ nhọn là 31.94, giá trị độ nhọn cao cho thấy phân phối lowest_yearly_earnings có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.13. Highest monthly earnings

Các đại lượng về xu thế trung tâm của biến highest_yearly_earnings
 Mean: 7081813.92
 Mode: [0.0]
 Median: 2600000.0

Các đại lượng về độ phân tán biến highest_yearly_earnings
 Khoảng biến thiên (Range): 163400000.0
 Phương sai (Variance): 190358246948510.34
 Độ lệch chuẩn (Standard deviation): 13797037.615

Các đại lượng về hình dáng phân phối biến highest_yearly_earnings
 Độ lệch: 4.79
 Độ nhọn: 31.87

Nhận xét:

- Biến highest_yearly_earnings có giá trị trung bình là 7081813.92, trung vị là 2600000.0 và giá trị thường xuất hiện nhất là 0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua độ biến thiên, giá trị phương sai và độ lệch chuẩn lớn.
- Độ lệch là 4.79, giá trị độ lệch dương cho thấy phân phối highest_yearly_earnings bị lệch sang phải, có nghĩa là có nhiều đối tượng có số lượng highest_yearly_earnings thấp hơn giá trị trung bình, và một số ít đối tượng có số lượng highest_yearly_earnings cao hơn nhiều so với giá trị trung bình; Độ nhọn là 31.87, giá trị độ nhọn cao cho thấy phân phối highest_yearly_earnings có đỉnh nhọn hơn so với phân phối chuẩn.

3.1.4.2.14. Created year

Các đại lượng về xu thế trung tâm của biến created_year
 Mean: 2012.63

Mode: [2014.0]

Median: nan

Các đại lượng về độ phân tán biến created_year

Khoảng biến thiên (Range): nan

Phương sai (Variance): 20.363

Độ lệch chuẩn (Standard deviation): 4.513

Các đại lượng về hình dáng phân phối biến created_year

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến created_year có giá trị trung bình là 2013 và giá trị thường xuất hiện nhất là 2014.
- Dữ liệu có mức độ phân tán trung bình, thể hiện qua giá trị phương sai và độ lệch chuẩn, các kênh được tạo tương đối đồng đều giữa các năm.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến created_year.

3.1.4.2.15. Created date

Các đại lượng về xu thế trung tâm của biến created_date

Mean: 15.746

Mode: [9.0]

Median: nan

Các đại lượng về độ phân tán biến created_date

Khoảng biến thiên (Range): nan

Phương sai (Variance): 77.045

Độ lệch chuẩn (Standard deviation): 8.778

Các đại lượng về hình dáng phân phối biến created_date

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến created_date có giá trị trung bình là 16 và giá trị thường xuất hiện nhất là 9.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn, các kênh được tạo có vẻ không đồng đều giữa các ngày trong tháng.

- Không có đủ thông tin để đánh giá hình dạng phân phối của biến created_date.

3.1.4.2.16. Gross tertiary education enrollment (%)

Các đại lượng về xu thế trung tâm của biến Gross tertiary education enrollment (%)

Mean: 63.628

Mode: [88.2]

Median: nan

Các đại lượng về độ phân tán biến Gross tertiary education enrollment (%)

Khoảng biến thiên (Range): nan

Phương sai (Variance): 681.57

Độ lệch chuẩn (Standard deviation): 26.107

Các đại lượng về hình dáng phân phối biến Gross tertiary education enrollment (%)

Độ lệch: nan

Độ nhọn: nan

Nhận xét:

- Biến Gross tertiary education enrollment (%) có giá trị trung bình là 63.628 và giá trị thường xuất hiện nhất là 88.2.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Gross tertiary education enrollment (%).

3.1.4.2.17. Unemployment rate

Các đại lượng về xu thế trung tâm của biến Unemployment rate

Mean: 9.279

Mode: [14.7]

Median: nan

Các đại lượng về độ phân tán biến Unemployment rate

Khoảng biến thiên (Range): nan

Phương sai (Variance): 23.896

Độ lệch chuẩn (Standard deviation): 4.888

Các đại lượng về hình dáng phân phối biến Unemployment rate

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến Unemployment rate có giá trị trung bình là 9.279 và giá trị thường xuất hiện nhất là 14.7.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Unemployment rate.

3.1.4.2.18. Population

Các đại lượng về xu thế trung tâm của biến Population

Mean: 430387266.752

Mode: [328239523.0]

Median: nan

Các đại lượng về độ phân tán biến Population

Khoảng biến thiên (Range): nan

Phương sai (Variance): 2.2353484959032054e+17

Độ lệch chuẩn (Standard deviation): 472794722.465

Các đại lượng về hình dáng phân phối biến Population

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến Population có giá trị trung bình là 430387266.752 và giá trị thường xuất hiện nhất là 328239523.0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Population.

3.1.4.2.19. Urban population

Các đại lượng về xu thế trung tâm của biến Urban_population

Mean: 224214981.632

Mode: [270663028.0]

Median: nan

Các đại lượng về độ phân tán biến Urban_population

Khoảng biến thiên (Range): nan

Phương sai (Variance): 2.392818018201603e+16

Độ lệch chuẩn (Standard deviation): 154687362.709

Các đại lượng về hình dáng phân phối biến Urban_population

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến Urban_population có giá trị trung bình là 224214981.632 và giá trị thường xuất hiện nhất là 270663028.0.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn lớn.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Urban_population.

3.1.4.2.20. Latitude

Các đại lượng về xu thế trung tâm của biến Latitude

Mean: 26.633

Mode: [37.09]

Median: nan

Các đại lượng về độ phân tán biến Latitude

Khoảng biến thiên (Range): nan

Phương sai (Variance): 422.736

Độ lệch chuẩn (Standard deviation): 20.561

Các đại lượng về hình dáng phân phối biến Latitude

Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến Latitude có giá trị trung bình là 26.633 và giá trị thường xuất hiện nhất là 37.09.
- Dữ liệu có mức độ phân tán trung bình, thể hiện qua giá trị phương sai và độ lệch chuẩn.

- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Latitude.

3.1.4.2.21. Longitude

Các đại lượng về xu thế trung tâm của biến Longitude

Mean: -14.128

Mode: [-95.713]

Median: nan

Các đại lượng về độ phân tán biến Longitude

Khoảng biến thiên (Range): nan

Phương sai (Variance): 7184.395

Độ lệch chuẩn (Standard deviation): 84.761

Các đại lượng về hình dáng phân phối biến Longitude

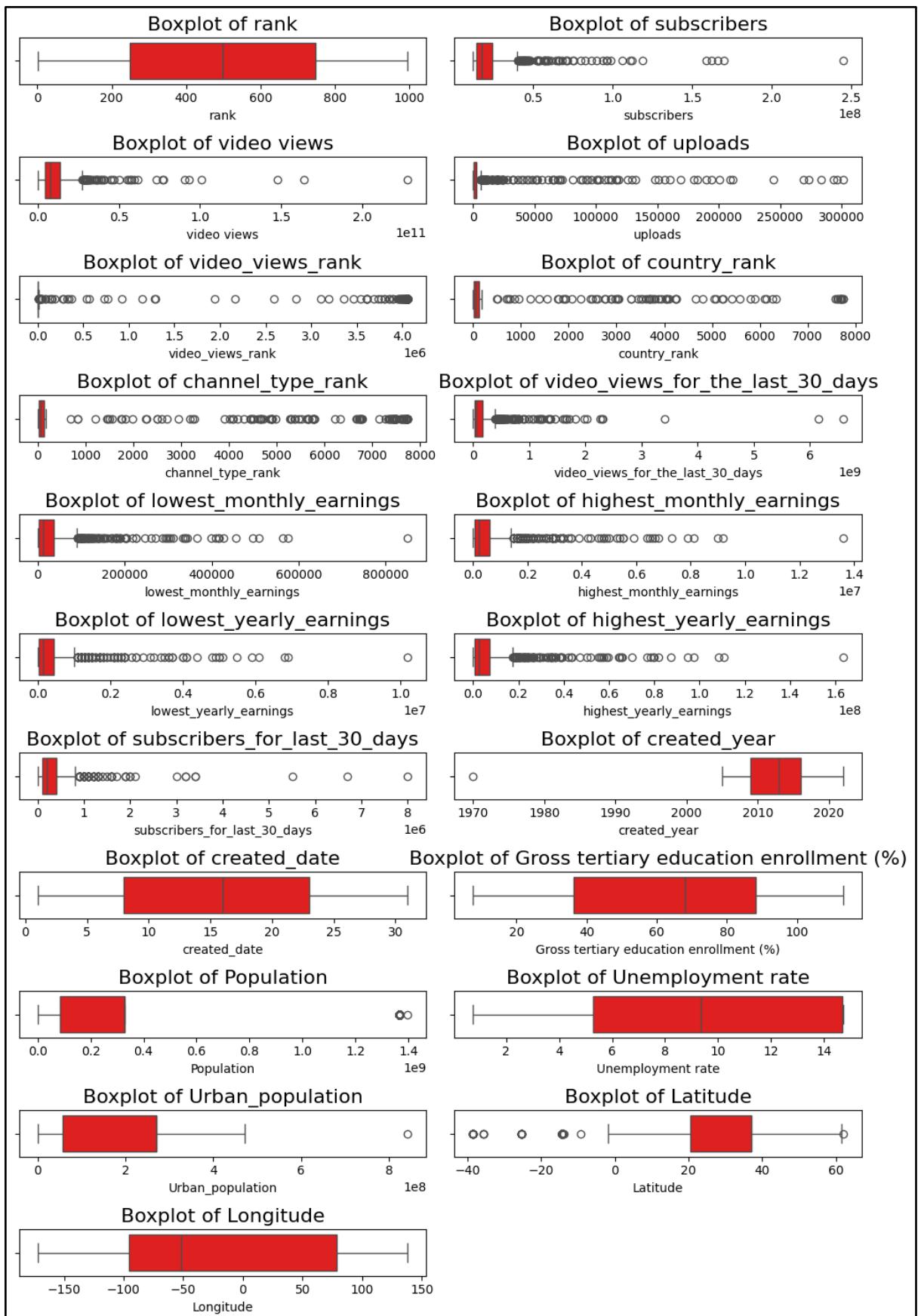
Độ lệch: nan

Độ nhọn: nan

Nhân xét:

- Biến Longitude có giá trị trung bình là -14.128 và giá trị thường xuất hiện nhất là -95.713.
- Dữ liệu có mức độ phân tán rộng, thể hiện qua giá trị phương sai và độ lệch chuẩn cao.
- Không có đủ thông tin để đánh giá hình dạng phân phối của biến Longitude.

3.1.4.2.21. Đánh giá độ phân tán của các biến thông qua boxplot



Hình 10 - Biểu đồ boxplot của các biến numeric (trước tiền xử lý)

Dựa trên các boxplot, có thể thấy rằng biến Rank có dạng phân phối đối xứng.

Phần lớn các biến numeric còn lại có phân phối lệch, cụ thể:

- Lệch trái gồm các biến: 'Created Year', 'Unemployment Rate' và 'Gross Tertiary Education Enrollment (%)' có nhiều dữ liệu rơi vào phía phải của trực, nhưng có một số giá trị "cực thấp" kéo dài về phía trái. Trong trường hợp này, giá trị mean thường nhỏ hơn median.
- Lệch phải gồm biến: 'Subscribers', 'Video Views', 'Uploads', 'Lowest Monthly Earnings', 'Highest Monthly Earnings', 'Lowest Yearly Earnings', 'Highest Yearly Earnings', 'Video Views Rank', 'Country Rank', 'Channel Type Rank', 'Video Views For The Last 30 Days', 'Subscribers For Last 30 Days', 'Created Date', 'Population', 'Urban Population', 'Latitude', 'Longitude'. Có nghĩa là có nhiều dữ liệu rơi vào phía trái của trực, nhưng có một số giá trị "cực cao" kéo dài về phía phải. Trong trường hợp này, giá trị mean thường lớn hơn median.

Đồng thời trong số này, một số biến có xuất dữ liệu nhiễu (outliers): Subscribers, Video Views, Uploads, Video Views Rank, Country Rank, Channel Type Rank, Video Views For The Last 30 Days, Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Yearly Earnings, Highest Yearly Earnings, Subscribers For Last 30 Days, Created Year, Population, Urban Population và Latitude.

3.2. Giai đoạn 2: Chuyển đổi dữ liệu (Transform)

Trước khi đến **Giai đoạn 3: Nạp dữ liệu vào kho dữ liệu (Load)**, nhóm sẽ tiến hành **Chuyển đổi dữ liệu (Transform)** bằng ngôn ngữ Python.

- Nguồn: [BI_Data Preprocessing.ipynb](#)
- Bộ dữ liệu ban đầu:
- Bộ dữ liệu sau khi tiền xử lý:
- Các bảng dimension và fact:

3.2.1. Thăm dò dữ liệu

Đầu tiên, nhóm sẽ tiến hành thăm dò dữ liệu *Data Exploration* để nắm được những thông tin cơ bản như số lượng thuộc tính, bản ghi được ghi nhận trong bộ dữ liệu. Qua đó, giúp nhóm có thể xác định được quy mô cũng như các thông tin quan trọng của bộ dữ liệu mà nhóm sẽ cần quan tâm đến.

```

1 data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995 entries, 0 to 994
Data columns (total 28 columns):
 #   Column           Non-Null Count Dtype
 --- 
 0   rank             995 non-null    int64
 1   Youtuber         995 non-null    object
 2   subscribers      995 non-null    int64
 3   video views     995 non-null    float64
 4   category         949 non-null    object
 5   Title            995 non-null    object
 6   uploads          995 non-null    int64
 7   Country           873 non-null    object
 8   Abbreviation     873 non-null    object
 9   channel_type     965 non-null    object
 10  video_views_rank 994 non-null    float64
 11  country_rank     879 non-null    float64
 12  channel_type_rank 962 non-null    float64
 13  video_views_for_the_last_30_days 939 non-null    float64
 14  lowest_monthly_earnings 995 non-null    float64
 15  highest_monthly_earnings 995 non-null    float64
 16  lowest_yearly_earnings 995 non-null    float64
 17  highest_yearly_earnings 995 non-null    float64
 18  subscribers_for_last_30_days 658 non-null    float64
 19  created_year      990 non-null    float64
 20  created_month     990 non-null    object
 21  created_date      990 non-null    float64
 22  Gross tertiary education enrollment (%) 872 non-null    float64
 23  Population        872 non-null    float64
 24  Unemployment rate 872 non-null    float64
 25  Urban_population 872 non-null    float64
 26  Latitude          872 non-null    float64
 27  Longitude         872 non-null    float64
dtypes: float64(18), int64(3), object(7)
memory usage: 217.8+ KB

```

Hình 11 - Thông tin các thuộc tính trong bộ dữ liệu

Qua việc tìm hiểu sơ lược bộ dữ liệu, nhóm có thể thu được các thông tin cơ bản như sau: bộ dữ liệu gồm **28 thuộc tính** và **995 bản ghi** được ghi nhận.

Tiếp đến, chúng ta sẽ in ra 5 dòng dữ liệu đầu tiên và cuối cùng của bộ dữ liệu để hiểu rõ hơn về cấu trúc và nội dung của nó.

	rank	Youtuber	subscribers	video views	category	Title	uploads	Country	Abbreviation	channel_type	...
0	1	T-Series	245000000	2.280000e+11	Music	T-Series	20082	India	IN	Music	...
1	2	YouTube Movies	170000000	0.000000e+00	Film & Animation	youtubemovies	1	United States	US	Games	...
2	3	MrBeast	166000000	2.836884e+10	Entertainment	MrBeast	741	United States	US	Entertainment	...
3	4	Cocomelon - Nursery Rhymes	162000000	1.640000e+11	Education	Cocomelon - Nursery Rhymes	966	United States	US	Education	...
4	5	SET India	159000000	1.480000e+11	Shows	SET India	116536	India	IN	Entertainment	...

5 rows × 28 columns

Hình 12 - Kết quả 5 dòng dữ liệu đầu tiên

rank		Youtuber	subscribers	video views	category	Title	uploads	Country	Abbreviation	channel_type	...
990	991	Natan por Aíç	12300000	9.029610e+09	Sports	Natan por Aíç	1200	Brazil	BR	Entertainment	...
991	992	Free Fire India Official	12300000	1.674410e+09	People & Blogs	Free Fire India Official	1500	India	IN	Games	...
992	993	Panda	12300000	2.214684e+09	Nan	HybridPanda	2452	United Kingdom	GB	Games	...
993	994	RobTopGames	12300000	3.741235e+08	Gaming	RobTopGames	39	Sweden	SE	Games	...
994	995	Make Joke Of	12300000	2.129774e+09	Comedy	Make Joke Of	62	India	IN	Comedy	...

5 rows × 28 columns

Hình 13 - Kết quả 5 dòng dữ liệu cuối cùng

Ta nhận thấy, nội dung của các dòng dữ liệu ở phần đầu bộ dữ liệu đều đầy đủ và có định dạng đúng. Tuy nhiên, ở các dòng dữ liệu ở phần cuối ta lại phát hiện dữ liệu có sự sai lệch trong bộ dữ liệu. Ví dụ như cột **Youtuber** chứa các ký tự lạ “í” và “ç” không phải là ký tự hợp lệ trong tên **YouTuber**, những ký tự này có thể là do lỗi nhập dữ liệu hoặc do cách thức thu thập dữ liệu; cột **Category** có chứa dữ liệu thiếu. Ở nội dung tiếp theo, ta sẽ áp dụng một số phương pháp để xử lý những vấn đề này.

3.2.2. Làm sạch dữ liệu

3.2.2.1. Đổi tên và loại bỏ cột

Đầu tiên, nhóm sẽ đặt lại tên cho các thuộc tính bằng cách xóa bỏ dấu “_” và thay thế bằng khoảng trắng và viết hoa các từ, nhằm giúp bộ dữ liệu có tính thống nhất và dễ dàng truy cập.

```

1 data.columns = data.columns.str.replace('_', ' ')
2 data.columns = data.columns.str.title()
3 data.columns

Index(['Rank', 'Youtuber', 'Subscribers', 'Video Views', 'Category', 'Title',
       'Uploads', 'Country', 'Abbreviation', 'Channel Type',
       'Video Views Rank', 'Country Rank', 'Channel Type Rank',
       'Video Views For The Last 30 Days', 'Lowest Monthly Earnings',
       'Highest Monthly Earnings', 'Lowest Yearly Earnings',
       'Highest Yearly Earnings', 'Subscribers For Last 30 Days',
       'Created Year', 'Created Month', 'Created Date',
       'Gross Tertiary Education Enrollment (%)', 'Population',
       'Unemployment Rate', 'Urban Population', 'Latitude', 'Longitude'],
      dtype='object')

```

Hình 14 - Kết quả xử lý tên cột

Tiếp theo đó, nhóm sẽ loại bỏ các cột không cần thiết vì không nằm trong mục tiêu cần phân tích như **Title**, **Video Views For The Last 30 Days**, **Subscribers For Last 30 Days**, **Latitude**, **Longitude**.

```

1 data = data.drop(columns=["Title", "Video Views For The Last 30 Days",
2                               "Subscribers For Last 30 Days", "Latitude", "Longitude", "Abbreviation"])

```

Hình 15 - Loại bỏ các cột dư thừa

3.2.2.2. Loại bỏ ký tự lạ và các trùng lặp của cột Youtuber

Như đã được đề cập ở nội dung trước, ở cột **Youtuber** nhóm phát hiện có sự sai lệch trong dữ liệu. Do vậy, ở đây nhóm sẽ xử lý các dữ liệu sai lệch này bằng cách thay thế các ký tự lạ đó bằng khoảng trắng.

```

1 pattern = r'[^a-zA-Z0-9\s.,!?&\-]'
2 # Youtuber
3 data['Youtuber'] = data['Youtuber'].apply(lambda x: re.sub(pattern, '', x))
4 data['Youtuber'] = data['Youtuber'].str.strip()
4 data['Youtuber'] = data['Youtuber'].str.lstrip(' ')

```

```
1 data.tail()
```

	Rank	Youtuber	Subscribers	Video Views	Category	Uploads	Country	Channel Type	Video Views Rank	Country Rank	...
990	991	Natan por A	12300000	9.029610e+09	Sports	1200	Brazil	Entertainment	525.0	55.0	...
991	992	Free Fire India Official	12300000	1.674410e+09	People & Blogs	1500	India	Games	6141.0	125.0	...
992	993	Panda	12300000	2.214684e+09	NaN	2452	United Kingdom	Games	129005.0	867.0	...
993	994	RobTopGames	12300000	3.741235e+08	Gaming	39	Sweden	Games	35112.0	4.0	...
994	995	Make Joke Of	12300000	2.129774e+09	Comedy	62	India	Comedy	4568.0	125.0	...

5 rows × 22 columns

Hình 16 - Kết quả xử lý dữ liệu trong cột Youtuber

Sau khi đã xử lý thay thế các ký tự lạ trong cột **Youtuber** bằng khoảng trắng, nhóm cần tiến hành kiểm tra lại liệu các dòng dữ liệu có **Youtuber** chứa giá trị trống hay không, trong trường hợp tên **Youtuber** toàn chứa các ký tự lạ.

```

1 # check Youtuber column after removed characters
2 blanks_data_youtuber = data[data['Youtuber'] == '']
3 blanks_data_youtuber

```

Hình 17 - Kiểm tra các giá trị trống trong cột Youtuber

Cuối cùng, nhóm sẽ tiến hành kiểm tra lại xem có dòng dữ liệu bị trùng lặp không và thực hiện các bước xử lý tiếp theo. Tuy nhiên, bộ dữ liệu không chứa dòng dữ liệu trùng lặp nào, nhóm sẽ không cần xử lý ở bước này.

```

1 # Check for duplicate rows
2 duplicate_rows = data[data.duplicated()]
3
4 # Print the number of duplicate rows
5 print(f"Number of duplicates: {len(duplicate_rows)}")
6
7 # Print the duplicate rows
8 print(duplicate_rows)

```

Number of duplicates: 0

Hình 18 - Kiểm tra giá trị trùng lặp trong bộ dữ liệu

Rank	Youtuber	Subscribers	Video Views	Category	Uploads	Country	Channel Type	Video Views Rank	Country Rank	...
64	65	43200000	3.645873e+10	Film & Animation	1478	Russia	Education	26.0	2.0	...
161	162	30400000	1.799996e+10	Nan	532	United States	Entertainment	147.0	46.0	...
433	434	19000000	4.924054e+09	Education	60	United States	Entertainment	1399.0	116.0	...
447	448	18800000	9.594189e+09	Entertainment	530	Ukraine	Entertainment	483.0	2.0	...
561	562	16500000	2.440934e+09	Entertainment	421	United States	Entertainment	3778.0	137.0	...
606	607	15900000	1.845330e+09	People & Blogs	0	Russia	Games	4057944.0	3309.0	...
632	633	15500000	8.265130e+09	Shows	1021	Saudi Arabia	Film	619.0	4.0	...
700	701	14900000	4.390980e+05	People & Blogs	1	Russia	News	3609784.0	10.0	...
707	708	14800000	7.018015e+09	Howto & Style	2387	Ukraine	Howto	828.0	3.0	...
714	715	14700000	2.230986e+09	Entertainment	1385	Saudi Arabia	Entertainment	4276.0	5.0	...
752	753	14400000	6.001543e+08	Entertainment	364	South Korea	Entertainment	21132.0	14.0	...
784	785	14100000	3.920221e+09	Entertainment	65	Nan	Nan	3999155.0	Nan	...
795	796	14000000	7.719743e+09	Gaming	2210	Russia	Games	703.0	12.0	...
810	811	13900000	8.451755e+09	Nan	504	United States	People	600.0	161.0	...
866	867	13300000	6.482687e+09	People & Blogs	608	Russia	Entertainment	927.0	13.0	...
920	921	12900000	5.585085e+09	Shows	1255	Ukraine	Entertainment	1164.0	6.0	...

16 rows × 22 columns

Hình 19 - Kết quả bộ dữ liệu sau khi xử lý cột Youtuber và loại bỏ trùng lặp

Như đã dự đoán trước, sau khi xử lý các ký tự lạ trên, bộ dữ liệu xuất hiện những dòng dữ liệu mà cột **Youtuber** chứa giá trị rỗng. Vì thuộc tính **Youtuber** là một trong những thuộc tính quan trọng nhất để phân tích ở các nội dung sau, nhóm quyết định sẽ

xóa bỏ các dòng dữ liệu này.

```
1 # remove those blank rows
2 data = data.drop(blanks_data_youtuber.index)
3 data.reset_index(drop= True, inplace= True)
```

Hình 20 - Loại bỏ dòng trống trong cột Youtuber

Tiếp theo đó, nhóm sẽ tiến hành kiểm tra các kênh trùng lặp trong cột **Youtuber** (*giả định rằng mỗi một Youtuber chỉ có một kênh Youtube duy nhất*). Do vậy, nhóm phải đảm bảo rằng bộ dữ liệu không có sự trùng lặp nào giữa các tên **Youtuber**.

```
1 # print rows that youtubers has duplicate
2
3 youtubers = data['Youtuber'].value_counts()
4 duplicate_youtubers = youtubers[youtubers > 1].index.tolist()
5
6 for youtuber in duplicate_youtubers:
7     print(data[data['Youtuber'] == youtuber])
```

Hình 21 - Kiểm tra giá trị trùng lặp trong cột Youtuber

	Rank	Youtuber	Subscribers	Video Views	Category	Uploads	\
217	220	Beyonc	26000000	1.730896e+10	Music	240	
913	930	Beyonc	12800000	1.418561e+10	Music	168	

Hình 22 - Giá trị bị trùng lặp trong cột Youtuber

	Rank	Youtuber	Subscribers	Video Views	Category	Uploads	\
860	876	BIBO	13200000	1.148422e+09	Entertainment	192	
899	915	BIBO	12900000	3.178223e+09	Entertainment	193	

Hình 23 - Giá trị bị trùng lặp trong cột Youtuber

Sau khi kiểm tra, nhóm nhận ra có hai giá trị **Youtuber** trùng lặp là **Beyonc** và **BIBO**. Do vậy, nhóm sẽ chỉ giữ lại tên kênh mà có ngày tạo kênh gần nhất.

```
1 # remove rank colum has values 220 and 915
2
3 data = data[~data['Rank'].isin([220, 915])]
```

Hình 24 - Xử lý giá trị bị trùng lặp

3.2.2.3. Xử lý giá trị thiếu

Sau khi đã thực hiện các bước tiền xử lý cơ bản như đổi tên cột, loại bỏ cột không cần thiết, xóa bỏ trùng lặp. Bộ dữ liệu hiện tại còn **22 cột** và **977 dòng**, so với ban đầu là **28 cột** và **995 dòng**.

```

1 data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 977 entries, 0 to 978
Data columns (total 22 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   Rank             977 non-null    int64   
 1   Youtuber         977 non-null    object  
 2   Subscribers      977 non-null    int64   
 3   Video Views      977 non-null    float64 
 4   Category          933 non-null    object  
 5   Uploads           977 non-null    int64   
 6   Country            856 non-null    object  
 7   Channel Type      948 non-null    object  
 8   Video Views Rank  976 non-null    float64 
 9   Country Rank       862 non-null    float64 
 10  Channel Type Rank 945 non-null    float64 
 11  Lowest Monthly Earnings 977 non-null    float64 
 12  Highest Monthly Earnings 977 non-null    float64 
 13  Lowest Yearly Earnings 977 non-null    float64 
 14  Highest Yearly Earnings 977 non-null    float64 
 15  Created Year      972 non-null    float64 
 16  Created Month      972 non-null    object  
 17  Created Date        972 non-null    float64 
 18  Gross Tertiary Education Enrollment (%) 855 non-null    float64 
 19  Population          855 non-null    float64 
 20  Unemployment Rate  855 non-null    float64 
 21  Urban Population     855 non-null    float64 
dtypes: float64(14), int64(3), object(5)
memory usage: 175.6+ KB

```

Hình 25 - Thông tin các cột dữ liệu sau khi loại bỏ cột

Tiếp đến, nhóm sẽ kiểm tra giá trị thiếu có trong bộ dữ liệu và tiến hành xử lý chúng. Việc xử lý giá trị thiếu là một bước quan trọng trong quá trình tiền xử lý dữ liệu, nhằm đảm bảo tính chính xác và hiệu quả cho các phân tích tiếp theo.

1 data.isna().sum()	
Rank	0
Youtuber	0
Subscribers	0
Video Views	0
Category	44
Uploads	0
Country	121
Channel Type	29
Video Views Rank	1
Country Rank	115
Channel Type Rank	32
Lowest Monthly Earnings	0
Highest Monthly Earnings	0
Lowest Yearly Earnings	0
Highest Yearly Earnings	0
Created Year	5
Created Month	5
Created Date	5
Gross Tertiary Education Enrollment (%)	122
Population	122
Unemployment Rate	122
Urban Population	122
dtype: int64	

Hình 26 - Kiểm tra giá trị bị thiếu trong bộ dữ liệu

Sau khi kiểm tra giá trị thiếu, nhóm nhận thấy các cột **Category**, **Country**, **Channel Type**, **Video Views Rank**, **Country Rank**, **Channel Type Rank**, **Gross Tertiary Education Enrollment (%)**, **Population**, **Unemployment Rate**, **Urban Population** đều chứa các giá trị thiếu. Do vậy, nhóm sẽ tiến hành xử lý các giá trị thiếu trong các thuộc tính này.

Ở đây, nhóm sẽ xử lý bằng cách chia ra các thuộc tính chứa giá trị thiếu thành hai nhóm thuộc tính theo kiểu dữ liệu gồm các thuộc tính có kiểu dữ liệu số và các thuộc tính có kiểu dữ liệu định danh.

- Nhóm các thuộc tính có kiểu dữ liệu là số: Video Views Rank, Country Rank, Channel Type Rank, Created Year, Created Month, Created Date, Gross Tertiary Education Enrollment (%), Population, Unemployment Rate, Urban Population.
- Nhóm các thuộc tính có kiểu dữ liệu là định danh: Category, Country, Channel Type.

Tiếp theo đó, nhóm sẽ xử lý giá trị thiếu của nhóm các thuộc tính có kiểu dữ liệu là số trước. Sau đó sẽ là nhóm các thuộc tính có kiểu dữ liệu là định danh.

- ❖ Nhóm các thuộc tính có kiểu dữ liệu là số
- Video Views Rank, Country Rank, Channel Type Rank

```

1 # fill nan of Video Views Rank, Country Rank, Channel Type Rank with 0
2
3 # replace object datatype column missing values with 0
4 data['Video Views Rank'] = data['Video Views Rank'].fillna(0).astype(int)
5 data['Country Rank'] = data['Country Rank'].fillna(0).astype(int)
6 data['Channel Type Rank'] = data['Channel Type Rank'].fillna(0).astype(int)

```

Hình 27 - Dièn giá trị thiéu các cột Video Views Rank, Country Rank, Channel Type Rank

Các kênh Youtube mà có giá trị thiéu trong những thuộc tính trên, nhóm sẽ điền các giá trị thiéu đó bằng giá trị 0 thay vì xóa đi những dòng dữ liệu này.

- Created Year, Created Month, Created Date

Lowest Yearly Earnings	Highest Yearly Earnings	Created Year	Created Month	Created Date	Gross Tertiary Education Enrollment	Population (%)
0.0	0.0	NaN	NaN	NaN	NaN	NaN
0.0	0.0	NaN	NaN	NaN	NaN	NaN
0.0	0.0	NaN	NaN	NaN	113.1	25766605.0
0.0	0.0	NaN	NaN	NaN	51.3	212559417.0
0.0	0.0	NaN	NaN	NaN	88.2	328239523.0

Hình 28 - In giá trị thiéu trong cột Created Year

Sau khi quan sát các giá trị thiéu của cột **Created Year**, nhóm nhận thấy các giá trị thiéu của thuộc tính **Created Year**, đồng thời cũng là của **Created Month** và **Created Date**. Do vậy, ở đây nhóm chỉ cần xóa đi các dòng dữ liệu mà **Created Year**

chứa giá trị thiếu.

```
1 # remove null on Created Year column
2 data = data.dropna(subset=['Created Year'])
3
4 # change type float64 to int64
5 data['Created Year'] = data['Created Year'].astype('int64')
```

Hình 29 - Xóa giá trị thiếu trong cột Created Year

- Gross Tertiary Education Enrollment (%), Population, Unemployment Rate, Urban Population

Nhóm sẽ xử lý các giá trị trên bằng cách thay thế bằng giá trị trung bình. Sử dụng giá trị trung bình của các bản ghi còn lại trong cùng cột để thay thế cho các giá trị thiếu. Do vậy, đầu tiên nhóm sẽ tính toán giá trị trung bình của từng cột như sau.

```
1 # data without null
2 data_not_null = data.dropna(subset= ['Gross Tertiary Education Enrollment (%)',
3                                         'Unemployment Rate', 'Population', 'Urban Population'])
4
5 # Mean Education Enrollment
6 mean_edu = round(data_not_null['Gross Tertiary Education Enrollment (%)'].mean(),1)
7 print('Mean Gross Tertiary Education Enrollment:', mean_edu)
8
9 # Mean Unemployment Rate
10 mean_unemp = round(data_not_null['Unemployment Rate'].mean(),1)
11 print('Mean Unemployment Rate:', mean_unemp)
12
13 # Mean Population
14 mean_popu = round(data_not_null['Population'].mean(),0)
15 print('Mean Population:', mean_popu)
16
17 # Mean Urban Population
18 mean_urban = round(data_not_null['Urban Population'].mean(),0)
19 print('Mean Urban Population:', mean_urban)

Mean Gross Tertiary Education Enrollment: 63.2
Mean Unemployment Rate: 9.3
Mean Population: 436693609.0
Mean Urban Population: 226424465.0
```

Hình 30 - Tính giá trị trung bình trong các cột Gross Tertiary Education Enrollment (%), Population, Unemployment Rate, Urban Population

Sau khi đã tính toán giá trị trung bình, nhóm sẽ thay thế giá trị thiếu bằng các giá trị vừa tính toán được.

```
1 # replace object datatype column missing values with mean
2 data['Gross Tertiary Education Enrollment (%)'] = data['Gross Tertiary Education Enrollment (%)'].fillna(mean_edu)
3 data['Unemployment Rate'] = data['Unemployment Rate'].fillna(mean_unemp)
4 data['Population'] = data['Population'].fillna(mean_popu)
5 data['Urban Population'] = data['Urban Population'].fillna(mean_urban)
```

Hình 31 - Xử lý giá trị thiếu trong các cột Gross Tertiary Education Enrollment (%), Population, Unemployment Rate, Urban Population

Sau khi xử lý các giá trị thiếu của nhóm các thuộc tính có kiểu dữ liệu là số, nhóm sẽ kiểm tra lại lần nữa xem còn giá trị thiếu nào hay không. Nhằm đảm bảo rằng, bộ dữ liệu được xử lý đầy đủ và chính xác.

1 data.isna().sum()	
Rank	0
Youtuber	0
Subscribers	0
Video Views	0
Category	44
Uploads	0
Country	119
Channel Type	26
Video Views Rank	0
Country Rank	0
Channel Type Rank	0
Lowest Monthly Earnings	0
Highest Monthly Earnings	0
Lowest Yearly Earnings	0
Highest Yearly Earnings	0
Created Year	0
Created Month	0
Created Date	0
Gross Tertiary Education Enrollment (%)	0
Population	0
Unemployment Rate	0
Urban Population	0
dtype: int64	

Hình 32 - Kiểm tra giá trị thiếu trong bộ dữ liệu

Sau khi quan sát, nhóm nhận thấy không còn thuộc tính nào có kiểu dữ liệu là số có chứa giá trị thiếu. Do vậy, nhóm sẽ tiếp tục xử lý các giá trị thiếu của nhóm các thuộc tính có kiểu dữ liệu là định danh.

- ❖ Nhóm các thuộc tính có kiểu dữ liệu là định danh
 - Country

```
1 # fill null of country with Unknown
2
3 data['Country'] = data['Country'].fillna('Unknown')
```

Hình 33 - Xử lý các giá trị bị thiếu trong cột Country

Ở đây, nhóm điền giá trị *Unknown* cho thuộc tính **Country** thay vì xóa bỏ các dòng

dữ liệu này thay vì xóa bỏ các dòng dữ liệu có giá trị thiếu cho phép duy trì tất cả các kênh trong phân tích, ngay cả khi thiếu thông tin về quốc gia, hay thực hiện các phân tích hoặc lọc chi tiết hơn sau này với các kênh có quốc gia là *Unknown*.

- Category, Channel Type

Đầu tiên, nhóm sẽ quan sát các giá trị duy nhất của hai thuộc tính **Category** và **Channel Type** để đưa cách xử lý các giá trị thiếu của hai thuộc tính này. Sau khi đã xem xét các giá trị như ở phía dưới, nhóm nhận thấy các giá trị của hai thuộc tính **Category - Channel Type** này có sự tương đồng với nhau như *Music - Music, Film & Animation - Film, Entertainment - Entertainment, Education - Education, Gaming - Games, New & Politics - News,...* Các giá trị của hai thuộc tính sẽ được thể hiện cụ thể và đầy đủ hơn ở bảng dưới.

S T T	Category	Channe l Type
1	Music	Music
2	Gaming	Games
3	Entertainment	Entertai nment
4	Education	Educati on
5	People & Blogs	People
6	Sports	Sports
7	Film & Animation	Film
8	News & Politics	News
9	Comedy	Comedy
1	Howto & Style	Howto

0		
1	Nonprofits & Activism	Nonprofit
1	Trailers	Autos
1	Science & Technology	Tech
1	Pets & Animals	Animals
1	Shows	
1	Trailers	
1	Movies	
1	Travel & Events	

Bảng 4 - Bảng so sánh các trị của hai thuộc tính Category và Channel Type

Từ bảng trên, nhóm đưa ra một số nhận xét như sau: số giá trị của thuộc tính **Category** thì nhiều hơn thuộc tính **Channel Type** và giữa hai thuộc tính có một số giá trị hoàn toàn giống nhau và một số thì tương tự nhau (đã được chỉ ra ở trên).

```

1 # Print unique values of Category and Channel Type
2 print('Category:', data['Category'].unique())
3 print('Channel Type:', data['Channel Type'].unique())

Category: ['Music' 'Film & Animation' 'Entertainment' 'Education' 'Shows' 'nan'
'People & Blogs' 'Gaming' 'Sports' 'Howto & Style' 'News & Politics'
'Comedy' 'Trailers' 'Nonprofits & Activism' 'Science & Technology'
'Movies' 'Pets & Animals' 'Autos & Vehicles' 'Travel & Events']
Channel Type: ['Music' 'Games' 'Entertainment' 'Education' 'People' 'Sports' 'Film'
'News' 'nan' 'Comedy' 'Howto' 'Nonprofit' 'Autos' 'Tech' 'Animals']

```

Hình 34 - In giá trị duy nhất trong cột Category và Channel Type

1 data[['Rank', 'Category', 'Channel Type']].head(15)			
	Rank	Category	Channel Type
0	1	Music	Music
1	2	Film & Animation	Games
2	3	Entertainment	Entertainment
3	4	Education	Education
4	5	Shows	Entertainment
5	6	Nan	Music
6	7	People & Blogs	Entertainment
7	8	Gaming	Entertainment
8	9	People & Blogs	People
9	10	Entertainment	Entertainment
10	11	Music	Music
11	12	Sports	Sports
12	13	Nan	Games
13	14	People & Blogs	Music
14	15	Film & Animation	Music

Hình 35 - Kết quả 15 dòng đầu của 3 cột Rank, Category và Channel Type

Do vậy, nhóm quyết định sẽ điền giá trị thiêu của cột **Category** bằng các giá trị của cột **Channel Type** trên cùng một dòng dữ liệu và xử lý tương tự như với các giá trị thiêu của **Channel Type**, là thay thế những giá trị này bằng giá trị của **Category** cũng trên một dòng dữ liệu.

Ví dụ 1, như ảnh trên ở dòng dữ liệu có **Rank** là 6, thì **Category** có chứa giá trị thiêu và **Channel Type** mang giá trị là *Music*. Vì vậy, nhóm sẽ thay thế giá trị thiêu ở thuộc tính **Category** có **Rank** là 6 bằng giá trị của **Channel Type** là *Music* mà cũng có **Rank** là 6.

Ví dụ 2, như ảnh trên ở dòng dữ liệu có **Rank** là 13, thì **Category** có chứa giá trị thiêu và **Channel Type** mang giá trị là *Games*. Vì vậy, nhóm sẽ thay thế giá trị thiêu ở thuộc tính **Category** có **Rank** là 13 bằng giá trị của **Channel Type** là *Games* mà cũng có **Rank** là 13.

```

1 # Fill same values on the same rows of Channel Type to Category if Category is null
2 data['Category'] = data.apply(lambda row: row['Channel Type'] if pd.isna(row['Category'])
3                                     else row['Category'], axis=1)
4
5 # Fill same values on the same rows of Category to Channel Type if Channel is null
6 data['Channel Type'] = data.apply(lambda row: row['Category'] if pd.isna(row['Channel Type'])
7                                     else row['Channel Type'], axis=1)

```

Hình 36 - Xử lý giá trị thiếu trong cột Category và Channel Type

Tuy nhiên khác với *Ví dụ 1*, ở *Ví dụ 2* này, giá trị được ánh xạ là *Games* khác với giá trị ban đầu của **Category** là *Gaming*. Do vậy, sau khi ánh xạ, nhóm cần thực hiện thêm một bước nữa là thay thế các giá trị được ánh xạ về giá trị gốc bởi các giá trị của **Category - Channel Type** chỉ mang tính tương đồng, không giống nhau hoàn toàn, như đã được nêu cụ thể trong *Bảng so sánh các giá trị của hai thuộc tính Category và Channel Type*.

```

1 # Print unique values of Category and Channel Type
2 # after mapping
3 print('Category after mapping:', data['Category'].unique())
4 print('Channel Type after mapping:', data['Channel Type'].unique())

Category after mapping: ['Music' 'Film & Animation' 'Entertainment' 'Education' 'Shows'
'People & Blogs' 'Gaming' 'Sports' 'Games' 'Howto & Style'
'News & Politics' 'Comedy' 'Trailers' 'Nonprofits & Activism' nan
'Science & Technology' 'Movies' 'People' 'Pets & Animals'
'Autos & Vehicles' 'Howto' 'Tech' 'Film' 'Travel & Events']
Channel Type after mapping: ['Music' 'Games' 'Entertainment' 'Education' 'People' 'Sports' 'Film'
'News' 'Howto & Style' 'Comedy' 'Howto' 'Nonprofit' nan 'Autos' 'Tech'
'Gaming' 'Film & Animation' 'People & Blogs' 'Animals' 'Shows'
'Pets & Animals' 'Science & Technology']

```

Hình 37 - In giá trị duy nhất trong cột Category và Channel Type sau khi điền dữ liệu bị thiếu

Sau khi ánh xạ, nhóm nhận thấy các giá trị của **Category** và **Channel Type** đã thay đổi so với ban đầu. Cụ thể, **Category** đã có thêm các giá trị như là *Games, People, Howto, Tech, Film*; **Channel Type** cũng phát sinh thêm các giá trị như *Howto & Style, Gaming, Film & Animation, People & Blogs, Shows, Pets & Animals, Science & Technology*.

Do vậy, nhóm cần tiến hành chuyển đổi các giá trị này về giá trị gốc để đảm bảo bộ dữ liệu vẫn chính xác và thống nhất như ban đầu, như trong thuộc tính **Category** nhóm sẽ chuyển đổi giá trị *Games* → *Gaming*. Các bước thực hiện xử lý chuyển đổi được trình bày cụ thể như hình sau.

```

1 ## Games => Gaming, People => People & Blogs, Howto => Howto & Style
2 # Tech => Science & Technology, Film => Film & Animation
3 # Replace values in 'Category' column
4 data['Category'] = data['Category'].replace({'Games': 'Gaming', 'People': 'People & Blogs',
5                                              'Howto': 'Howto & Style', 'Tech': 'Science & Technology',
6                                              'Film': 'Film & Animation'})
7
8
9

1 ## Howto & Style => Howto, Gaming => Games, Film & Animation => Film, People & Blogs => People
2 # Shows => Entertainment, Pet & Animals => Animals, Science & Technology => Tech
3 # Replace values in 'Channel Type' column
4 data['Channel Type'] = data['Channel Type'].replace({'Howto & Style': 'Howto', 'Gaming': 'Games',
5                                                       'Film & Animation': 'Film', 'People & Blogs': 'People', 'Shows': 'Entertainment',
6                                                       'Pet & Animals': 'Animals', 'Science & Technology': 'Tech'})
```

Hình 38 - Xử lý các giá trị trong cột **Category** và **Channel Type**

Sau khi đã thực hiện các bước điền giá trị thiếu bằng cách ánh xạ giá trị của cột **Category** sang cột **Channel Type** và ngược lại, cũng như chuyển đổi các giá trị như đã được trình bày ở phía trên. Cuối cùng, nhóm sẽ kiểm tra lại xem còn giá trị thiếu nào hay không.

1	data.isna().sum()
Rank	0
Youtuber	0
Subscribers	0
Video Views	0
Category	3
Uploads	0
Country	0
Channel Type	3
Video Views Rank	0
Country Rank	0
Channel Type Rank	0
Lowest Monthly Earnings	0
Highest Monthly Earnings	0
Lowest Yearly Earnings	0
Highest Yearly Earnings	0
Created Year	0
Created Month	0
Created Date	0
Gross Tertiary Education Enrollment (%)	0
Population	0
Unemployment Rate	0
Urban Population	0
dtype: int64	

Hình 39 - Kiểm tra giá trị bị thiếu sau khi xử lý giá trị thiếu ở tất cả các cột

Sau khi kiểm tra, nhóm nhận thấy cột **Category** và **Channel Type** vẫn còn chứa các giá trị thiếu, cụ thể là 3 giá trị. Do vậy, để có nhiều thông tin hơn nhóm sẽ tiến hành in ra 3 dòng dữ liệu này để quan sát.

```

1 # Print row that category and channel type have null values
2
3 data[(data['Category'].isna() | (data['Channel Type'].isna()))]
4

```

Rank	Youtuber	Subscribers	Video Views	Category	Uploads	Country	Channel Type
103	News	36300000	0	NaN	0	Unknown	NaN
446	Busy Fun Ltd	18800000	7762077012	NaN	2	Unknown	NaN
593	Live	16100000	0	NaN	0	Unknown	NaN

Hình 40 - Kết quả 3 dòng giá trị thiếu ở cột Category và Channel Type

Từ hình ảnh trên, nhóm đưa ra nhận xét như sau cả hai thuộc tính đều có số giá trị thiếu bằng nhau và đều có **Rank** là 103, 446, 593. Cách xử lý giá trị thiếu là ánh xạ các giá trị từ **Category** đến **Channel Type** và ngược lại. Tuy nhiên ở ba trường hợp này, cả hai thuộc tính **Category** và **Channel Type** đều chứa giá trị thiếu, nhóm không thể sử dụng phương pháp ánh xạ trên và cũng như là không có cơ sở nào để điền các giá trị thiếu này.

Do vậy, nhóm sẽ chọn cách giải quyết là xóa đi các dòng dữ liệu này khỏi bộ dữ liệu.

```

1 # Remove row that rank is 103 or 446 or 593
2 data = data.drop(data[data['Rank'] == 103].index)
3 data = data.drop(data[data['Rank'] == 446].index)
4 data = data.drop(data[data['Rank'] == 593].index)

```

Hình 41 - Xử lý 3 giá trị thiếu trong cột Category và Channel Type

Như đã được nêu rõ ở *Bảng so sánh các giá trị của hai thuộc tính Category và Channel Type*, nhóm nhận thấy thuộc tính **Channel Type** có số giá trị nhiều hơn so với **Category** và cũng tương đồng về mặt ý nghĩa.Thêm vào đó, để bộ dữ liệu mang tính thống nhất và chặt chẽ hơn để phân tích ở các nội dung sau, nhóm sẽ thay đổi thuộc tính **Channel Type** → **Category**, **Category** → **SubCategory**.

```

1 # Change name of Channel Type to Category, Category to Subcategory
2 data = data.rename(columns={'Channel Type': 'Category', 'Category': 'SubCategory'})
3
4 # Print the updated DataFrame
5 print(data.columns)

Index(['Rank', 'Youtuber', 'Subscribers', 'Video Views', 'SubCategory',
       'Uploads', 'Country', 'Category', 'Video Views Rank', 'Country Rank',
       'Channel Type Rank', 'Lowest Monthly Earnings',
       'Highest Monthly Earnings', 'Lowest Yearly Earnings',
       'Highest Yearly Earnings', 'Created Year', 'Created Month',
       'Created Date', 'Gross Tertiary Education Enrollment (%)', 'Population',
       'Unemployment Rate', 'Urban Population'],
      dtype='object')

```

Hình 42 - Kết quả 15 dòng đầu của 3 cột Rank, Category và Channel Type

3.2.2.4. Xử lý nhiễu

Sau khi quan sát chi tiết bộ dữ liệu, nhóm nhận thấy ở các cột **Video Views**, **Uploads**, **Lowest Monthly Earnings**, **Highest Monthly Earnings**, **Lowest Year Earnings**, **Highest Year Earnings**, **Created Year** đều chứa các giá trị bất thường. Do vậy, nhóm sẽ tiến hành tìm hiểu và xử lý các giá trị trong các thuộc tính này.

- **Created Year**

Đầu tiên, nhóm sẽ in ra các giá trị có thể có của **Created Year** thì nhóm phát hiện ra rằng có kênh youtube được tạo vào năm 1970. Tuy nhiên, Youtube được thành lập vào năm 2005, kênh mà được tạo vào năm 1970 được xem là nhiễu. Nhóm cần in dòng dữ liệu này để quan sát chi tiết hơn.

```

1 # Check unique values of Created Year
2 data['Created Year'].unique()

array([2006, 2012, 2015, 2010, 2016, 2018, 2014, 2007, 2020, 2008, 2005,
       2011, 2013, 2009, 2017, 2021, 1970, 2019, 2022])

```

Hình 43 - Kiểm tra giá trị duy nhất trong cột Created Year

```

1 # Check Created Year is 1970
2 data[['Rank', 'Created Date', 'Created Month', 'Created Year']][data['Created Year'] == 1970]
3 # this youtuber was on Jan 1, 1970

```

Rank	Created Date	Created Month	Created Year	
94	102	1	Jan	1970

Hình 44 - Kiểm tra dòng có Created Year là 1970

Như ảnh trên, nhóm quan sát được rằng kênh youtube này được tạo vào ngày 01 tháng 01 năm 1970. Vì vậy, ở đây nhóm sẽ thay thế năm tạo kênh từ 1970 sang 2006.

```

1 # replace the year 1970 with 2006 because Youtube was created in 2005
2 data[data['Rank'] == 102] = data[data['Rank'] == 102].replace(1970, 2006)
3 data[['Rank', 'Youtuber', 'Created Date', 'Created Month', 'Created Year']][data['Rank'] == 102]

```

Rank	Youtuber	Created Date	Created Month	Created Year
94	102	YouTube	1	Jan 2006

Hình 45 - Xử lý dữ liệu có Created Year là 1970

- Video Views, Uploads

Sau khi quan sát chi tiết bộ dữ liệu, nhóm nhận thấy rằng có một số kênh youtube có **Video Views** hoặc **Uploads** mang giá trị 0.

Rank	Youtuber	Subscribers	Video Views	SubCategory	Uploads	Country	Category
1	2	YouTube Movies	170000000	0	Film & Animation	1	United States Games
5	6	Music	119000000	0	Music	0	Unknown Music
12	13	Gaming	93600000	0	Gaming	0	Unknown Games
18	19	Sports	75000000	0	Entertainment	3	United States Entertainment
171	174	Popular on YouTube	29300000	0	Education	3	Unknown Education
358	361	Minecraft - Topic	20900000	0	Gaming	0	Unknown Games

6 rows × 22 columns

Hình 46 - In các giá trị 0 trong cột Video Views và Uploads

Đây là một điều bất hợp lý, vì bộ dữ liệu này là tập hợp các kênh youtube có thứ hạng cao trên toàn cầu, có số lượt xem và số lượt đăng ký đều rất cao. Do vậy, những trường hợp mà có **Video Views**, **Uploads** là 0 thì được xem là nhiều cần loại bỏ khỏi bộ dữ liệu.

```

1 # Remove zero data Video Views
2 data = data.drop(zero_views.index)
3 data.reset_index(drop= True, inplace= True)

```

Hình 47 - Xử lý giá trị 0 trong cột Video Views

```

1 # Remove zero data Uploads
2 data = data.drop(zero_uploads.index)
3 data.reset_index(drop= True, inplace= True)

```

Hình 48 - Xử lý giá trị 0 trong cột Uploads

Bên cạnh trường hợp các kênh mà có **Video Views, Uploads** là 0 và đã được xử lý ở phía trên, tương tự như vậy nhóm cho rằng các kênh mà có **Uploads** bé hơn 10 (tức là các kênh này chỉ có số video đăng tải trên Youtube ít hơn 10 video) cũng là các dữ liệu nhiễu và cần tiến hành xử lý.

```

1 # Check Uploads < 10
2 uploadslessthan10 = data[data['Uploads'] < 10]
3 uploadslessthan10.head(10)

```

Rank	Youtuber	Subscribers	Video Views	SubCategory	Uploads	Country	Category	Video Views Rank
11	15	Goldmines	86900000	24118230580	Film & Animation	1	Unknown	Music 4056562
13	17	5-Minute Crafts	80100000	26236790209	Howto & Style	1	United Kingdom	Entertainment 4057901
44	49	Badabun	46800000	19398045702	Entertainment	1	Unknown	Music 4047729

Hình 49 - In 10 dòng đầu tiên có số lượng đăng tải ít hơn 10

Ví dụ, kênh **Youtube Goldmines** có thứ hạng là 15, có hơn 24 tỷ lượt xem và sở hữu hơn 80 triệu lượt đăng ký. Tuy nhiên, theo thống kê kênh này chỉ có *duy nhất một* video được đăng tải trên Youtube. Đây được xem là một điều bất hợp lý và cần được xử lý một cách cẩn thận để đảm bảo bộ dữ liệu mang tính chính xác.

Ở trường hợp này, nhóm sẽ không xử lý bằng cách xóa bỏ đi các nhiễu này mà thay vào đó nhóm sẽ thay thế những giá trị của **Uploads < 10** bằng cách *điền* các giá trị **Uploads** của các kênh youtube này tính đến thời điểm hiện tại.

- Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Year Earnings, Highest Year Earnings

	Rank	Lowest Monthly Earnings	Highest Monthly Earnings	Lowest Yearly Earnings	Highest Yearly Earnings	
11	15	0.0	0.07	0.05	0.86	
13	17	0.0	0.00	0.00	0.05	
19	24	0.0	0.00	0.00	0.00	
23	28	0.0	0.04	0.03	0.48	
47	52	0.0	0.00	0.00	0.05	
57	63	0.0	0.02	0.01	0.19	
68	76	0.0	0.02	0.02	0.24	
70	78	0.0	0.01	0.01	0.10	
92	100	0.0	0.00	0.00	0.00	
113	123	0.0	0.00	0.00	0.00	

Hình 50 - In giá trị 0 trong các cột Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Year Earnings, Highest Year Earnings

Cũng tương tự như trên, nhóm nhận thấy ở các thuộc tính như **Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Year Earnings, Highest Year Earnings** có chứa các giá trị 0. Có thể là trong quá trình thu thập dữ liệu, đây là các dữ liệu về thu nhập/doanh số của các kênh nhưng không được công khai, dẫn đến các dữ liệu này có giá trị 0. Điều này cũng có thể được xem là dễ hiểu vì các chính sách bảo mật và quy định pháp lý, một số kênh có thể có chính sách bảo mật nghiêm ngặt, hạn chế việc công khai dữ liệu thu nhập/doanh số hoặc có quy định pháp lý hạn chế việc công khai thông tin tài chính, bao gồm cả thu nhập/doanh số.

Đồng thời, nhóm cũng ghi nhận được tần suất xuất hiện của các giá trị 0 này lên đến 8%. Chính vì thế, nhóm sẽ thay thế các giá trị này bằng trung vị, để giảm thiểu sự mất mát thông tin và tính chính xác của bộ dữ liệu.

1 print('Median of Lowest Monthly Earnings:', data['Lowest Monthly Earnings'].median())
2 print('Median of Highest Monthly Earnings:', data['Highest Monthly Earnings'].median())
3 print('Median of Lowest Yearly Earnings:', data['Lowest Yearly Earnings'].median())
4 print('Median of Highest Yearly Earnings:', data['Highest Yearly Earnings'].median())
Median of Lowest Monthly Earnings: 14700.0 Median of Highest Monthly Earnings: 235400.0 Median of Lowest Yearly Earnings: 176500.0 Median of Highest Yearly Earnings: 2800000.0

Hình 51 - Tính trung vị các cột Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Year Earnings, Highest Year Earnings

```

1 data['Lowest Monthly Earnings'] = data['Lowest Monthly Earnings'].replace(0.0, data['Lowest Monthly Earnings'].median())
2 data['Highest Monthly Earnings'] = data['Highest Monthly Earnings'].replace(0.0, data['Highest Monthly Earnings'].median())
3 data['Lowest Yearly Earnings'] = data['Lowest Yearly Earnings'].replace(0.0, data['Lowest Yearly Earnings'].median())
4 data['Highest Yearly Earnings'] = data['Highest Yearly Earnings'].replace(0.0, data['Highest Yearly Earnings'].median())

```

Hình 52 - Xử lý giá trị bằng 0 trong các cột Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Year Earnings, Highest Year Earnings

3.2.3. Tạo thêm thuộc tính

```

1 data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 931 entries, 0 to 930
Data columns (total 22 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Rank             931 non-null    int64  
 1   Youtuber         931 non-null    object  
 2   Subscribers      931 non-null    int64  
 3   Video Views      931 non-null    int64  
 4   SubCategory       931 non-null    object  
 5   Uploads          931 non-null    int64  
 6   Country           931 non-null    object  
 7   Category          931 non-null    object  
 8   Video Views Rank 931 non-null    int64  
 9   Country Rank     931 non-null    int64  
 10  Channel Type Rank 931 non-null    int64  
 11  Lowest Monthly Earnings 931 non-null    float64 
 12  Highest Monthly Earnings 931 non-null    float64 
 13  Lowest Yearly Earnings 931 non-null    float64 
 14  Highest Yearly Earnings 931 non-null    float64 
 15  Created Year     931 non-null    int64  
 16  Created Month    931 non-null    object  
 17  Created Date     931 non-null    int64  
 18  Gross Tertiary Education Enrollment (%) 931 non-null    float64 
 19  Population        931 non-null    int64  
 20  Unemployment Rate 931 non-null    float64 
 21  Urban Population 931 non-null    int64  
dtypes: float64(6), int64(11), object(5)
memory usage: 160.1+ KB

```

Hình 53 - In thông tin các cột trong bộ dữ liệu

Sau khi đã thực hiện các bước tiền xử lý cho bộ dữ liệu, bộ dữ liệu còn **931 dòng** và **22 cột**, bao gồm các thuộc tính nổi bật như **Subscribers**, **Video Views**, **Category**, **SubCategory**, **Country**. Tuy nhiên, nhóm nhận thấy rằng các thuộc tính hiện tại vẫn chưa đầy đủ để nhóm có thể thực hiện các phân tích tổng quát như thống kê số lượng kênh youtube theo châu lục, thống kê số lượng kênh youtube đạt được nút kim cương/vàng/bạc, hay từ cách hoạt động của các kênh youtube như công ty (company), nhóm (group), cá nhân (individual). Do vậy, nhóm sẽ tạo thêm 3 thuộc tính **Region**,

Youtube Creator Awards, Type of Youtuber để giúp làm phong phú, đa dạng thêm cho bộ dữ liệu.

- Region

```
1 data['Country'].unique()  
49  
  
1 data['Country'].unique()  
  
array(['India', 'United States', 'Japan', 'Russia', 'South Korea',  
       'Unknown', 'United Kingdom', 'Canada', 'Brazil', 'Argentina',  
       'Chile', 'Cuba', 'El Salvador', 'Pakistan', 'Philippines',  
       'Thailand', 'Colombia', 'Barbados', 'Mexico',  
       'United Arab Emirates', 'Spain', 'Saudi Arabia', 'Indonesia',  
       'Turkey', 'Venezuela', 'Kuwait', 'Jordan', 'Netherlands',  
       'Australia', 'Italy', 'Germany', 'France', 'Sweden', 'Ukraine',  
       'Latvia', 'Switzerland', 'Vietnam', 'Malaysia', 'China', 'Iraq',  
       'Singapore', 'Egypt', 'Andorra', 'Ecuador', 'Morocco', 'Peru',  
       'Bangladesh', 'Finland', 'Samoa'], dtype=object)
```

Hình 54 - In giá trị duy nhất trong cột Country

Sau khi quan sát sơ bộ các giá trị của thuộc tính **Country**, nhóm nhận thấy các kênh youtube đến từ 48 quốc gia như *United States (Hoa Kỳ)*, *India (Ấn Độ)*, *Japan (Nhật Bản)*, *South Korea (Hàn Quốc)*.... Như vậy, sau khi đã biết tất cả các giá trị của **Country**, nhóm sẽ tiến hành tạo danh sách các châu lục tương ứng với các quốc gia đó. Ví dụ, *United States (Hoa Kỳ)* sẽ thuộc về *Châu Mỹ*, các quốc gia khác như *India (Ấn Độ)*, *Japan (Nhật Bản)*, *South Korea (Hàn Quốc)* sẽ thuộc về *Châu Á*.

Đầu tiên, nhóm sẽ tạo cột mới có tên **Region** với giá trị *Unknown* cho mọi dòng.

```
1 # Create a new column named Region  
2 data['Region'] = 'Unknown'
```

Hình 55 - Tạo cột Region

Tiếp theo đó, nhóm sẽ tạo danh sách tên các châu lục gồm các quốc gia tương ứng. Ví dụ, nếu dòng đó có **Country** mang giá trị là *Russia* thì **Region** sẽ thay đổi giá trị từ *Unknown* sang *Asia*. Trong trường hợp **Country** mang giá trị là *Unknown* (tức là kênh youtube không có thông tin về quốc gia) thì **Region** cũng sẽ có giá trị là *Unknown*.

```

4 # Set Region to 'Asia' for specific countries
5 asia_countries = ['India', 'Japan', 'Pakistan', 'Philippines', 'Thailand', 'Afghanistan', 'Bangladesh',
6                 'Indonesia', 'Turkey', 'China', 'Iraq', 'Egypt', 'Vietnam', 'Malaysia', 'Singapore', 'South Korea']
7 data.loc[data['Country'].isin(asia_countries), 'Region'] = 'Asia'
8
9 # Set Region to 'Europe' for specific countries
10 europe_countries = ['Russia', 'United Kingdom', 'Netherlands', 'Sweden', 'Spain', 'Andorra',
11                      'Latvia', 'Switzerland', 'Germany', 'France', 'Italy', 'Finland', 'Ukraine']
12 data.loc[data['Country'].isin(europe_countries), 'Region'] = 'Europe'
13
14 # Set Region to 'America' for specific countries
15 america_countries = ['United States', 'Canada', 'Mexico', 'Barbados',
16                      'Brazil', 'Argentina', 'Chile', 'Cuba', 'El Salvador',
17                      'Colombia', 'Venezuela', 'Ecuador', 'Peru', 'Bolivia',
18                      ]
19 data.loc[data['Country'].isin(america_countries), 'Region'] = 'America'

```

Hình 56 - Điene giá trị vào cột Region

```

21 # Set Region to 'Africa' for specific countries
22 africa_countries = ['Morocco', 'United Arab Emirates', 'Kuwait', 'Jordan', 'Saudi Arabia']
23 data.loc[data['Country'].isin(africa_countries), 'Region'] = 'Africa'
24
25 # Set Region to 'Oceania' for specific countries
26 oceania_countries = ['Australia', 'Samoa']
27 data.loc[data['Country'].isin(oceania_countries), 'Region'] = 'Oceania'

```

Hình 57 - Điene giá trị vào cột Region

Sau khi đã hoàn tất việc tạo cột **Region**, nhóm ghi nhận được có 5 châu lục là *Asia* (*Châu Á*), *Europe* (*Châu Âu*), *America* (*Châu Mỹ*), *Africa* (*Châu Phi*), và cuối cùng là *Oceania* (*Châu Úc*).

```

1 print(data['Region'].unique())
['Asia' 'America' 'Europe' 'Unknown' 'Africa' 'Oceania']

```

Hình 58 - In giá trị duy nhất trong cột Region

```

29 # Print the updated DataFrame
30 data[['Rank', 'Country', 'Region']].head(10)

```

Rank	Country	Region	
0	1 India	Asia	
1	3 United States	America	
2	4 United States	America	
3	5 India	Asia	
4	7 United States	America	
5	8 Japan	Asia	
6	9 Russia	Europe	
7	10 United States	America	
8	11 India	Asia	
9	12 United States	America	

Hình 59 - In 10 dòng đầu tiên của các cột Rank, Country và Region

- Youtube Creator Awards

Nhóm nhận thấy, ngoài **Rank** (thứ hạng) thì việc có thêm một thuộc tính để phân loại các kênh youtube là điều cần thiết để nhóm có thể thực hiện thêm các phân tích tổng quát khác. Do vậy, thuộc tính **Youtube Creator Awards/Youtube Play Button**

- Phần thưởng cho Nhà sáng tạo khi đạt cột mốc về số lượng lượt đăng ký là một trong các thuộc tính quan trọng mà nhóm đặc biệt quan tâm đến.

Dựa trên cách phân loại kênh chính thức từ Youtube thì các kênh mà có số lượt đăng ký trên 100 triệu sẽ được nhận nút kim cương đỏ (*Red Diamond*), từ 10 triệu đến dưới 100 triệu lượt đăng ký sẽ được nhận nút kim cương (*Diamond*), từ 1 triệu đến dưới 10 triệu lượt đăng ký sẽ được nhận nút vàng (*Gold*), từ 100 nghìn đến dưới 1 triệu sẽ được nhận nút bạc.

```

1 data['Youtube Creator Awards'] = ''
2 for index, row in data.iterrows():
3     if row['Subscribers'] >= 100000000:
4         data.at[index, 'Youtube Creator Awards'] = 'Red Diamond'
5     elif 10000000 <= row['Subscribers'] < 100000000:
6         data.at[index, 'Youtube Creator Awards'] = 'Diamond'
7     elif 1000000 <= row['Subscribers'] < 1000000:
8         data.at[index, 'Youtube Creator Awards'] = 'Gold'
9     elif 100000 <= row['Subscribers'] < 1000000:
10        data.at[index, 'Youtube Creator Awards'] = 'Silver'
11    else:
12        data.at[index, 'Youtube Creator Awards'] = 'Other'

```

Hình 60 - Tạo cột Youtube Creator Awards

Sau khi đã tạo thuộc tính **Youtube Creator Awards** thỏa các điều kiện trên thì nhóm nhận thấy rằng tất cả các kênh youtube trong bộ dữ liệu đều sở hữu nút kim cương (*Red Diamond*) và nút kim cương (*Diamond*).

Tuy nhiên, để đảm bảo tính chính xác nhóm sẽ tiến hành kiểm tra lại bằng cách in ra số lượt đăng ký thấp nhất và cao nhất. Nhóm nhận thấy rằng, kênh mà có lượt đăng ký ít nhất là *12.300.000* và nhiều nhất là *245.000.000*, do vậy thuộc tính này chỉ gồm 2 giá trị là *Red Diamond* và *Diamond*. Đây cũng được xem là điều hợp lý khi mà các kênh trong bộ dữ liệu đều là các kênh sở hữu lượt đăng ký và lượt xem rất cao.

```

1 print(data['Youtube Creator Awards'].unique())
[ 'Red Diamond' 'Diamond' ]

```

Hình 61 - In giá trị duy nhất trong cột Youtube Creator Awards

```

1 # Min value of Subscribers
2 print('Minimun subcribers:', data['Subscribers'].min())
3
4 # Max value of Subscribers
5 print('Maximun subcribers:', data['Subscribers'].max())

```

```

Minimun subcribers: 12300000
Maximun subcribers: 245000000

```

Hình 62 - In giá trị lớn nhất và nhỏ nhất trong cột Subscribers

- Type of Youtuber

Thuộc tính cuối cùng mà nhóm muốn thêm vào bộ dữ liệu đó chính là **TypeofYoutuber**, thuộc tính này giúp phân loại các kênh youtube dựa trên loại hình hoạt động của họ, cho phép nhóm phân tích dữ liệu chi tiết hơn.

Ví dụ:

- Group: Nhóm nhạc, kênh hài tập thể, kênh giáo dục do nhiều người dẫn dắt.
- Individual: Kênh cá nhân của ca sĩ, diễn viên, streamer.
- Company: Kênh chính thức của doanh nghiệp, tổ chức, thương hiệu.

```
1 gys['TypeofYoutuber'].unique()  
array(['Company', 'Individual', 'Group'], dtype=object)
```

Hình 63 - In giá trị duy nhất trong cột TypeofYoutuber

```
1 gys[['Rank', 'Youtuber', 'TypeofYoutuber']].head(10)
```

	Rank	Youtuber	TypeofYoutuber
0	1	T-Series	Company
1	3	MrBeast	Individual
2	4	Cocomelon - Nursery Rhymes	Company
3	5	SET India	Company
4	7	Kids Diana Show	Company
5	8	PewDiePie	Individual
6	9	Like Nastya	Individual
7	10	Vlad and Niki	Group
8	11	Zee Music Company	Company
9	12	WWE	Company

Hình 64 - In 10 dòng đầu tiên trong các cột Rank, Youtuber và TypeofYoutuber

Kết luận, sau khi đã thực hiện các bước tiền xử lý như làm sạch dữ liệu gồm xử lý các giá trị thiếu, trùng lặp, nhiễu và thêm các thuộc tính khác như **Region**, **Type of Youtuber**, **Youtube Creator Award** thì bộ dữ liệu hiện tại có tên là **finaleGYS.xlsx**.

```

1 gys.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 930 entries, 0 to 929
Data columns (total 25 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   Rank             930 non-null   int64
 1   Youtuber         930 non-null   object
 2   Subscribers      930 non-null   int64
 3   Video Views      930 non-null   int64
 4   Uploads          930 non-null   int64
 5   Country          930 non-null   object
 6   Video Views Rank 930 non-null   int64
 7   Country Rank     930 non-null   int64
 8   Channel Type Rank 930 non-null   int64
 9   Lowest Monthly Earnings 930 non-null   float64
 10  Highest Monthly Earnings 930 non-null   float64
 11  Lowest Yearly Earnings 930 non-null   float64
 12  Highest Yearly Earnings 930 non-null   float64
 13  Created Year     930 non-null   int64
 14  Created Month    930 non-null   object
 15  Created Date     930 non-null   int64
 16  Gross Tertiary Education Enrollment (%) 930 non-null   float64
 17  Population        930 non-null   int64
 18  Unemployment Rate 930 non-null   float64
 19  Urban Population 930 non-null   int64
 20  Region            930 non-null   object
 21  Youtube Creator Awards 930 non-null   object
 22  Category          930 non-null   object
 23  SubCategory       930 non-null   object
 24  TypeofYoutuber   930 non-null   object
dtypes: float64(6), int64(11), object(8)
memory usage: 181.8+ KB

```

Hình 65 - In thông tin các cột trong bộ dữ liệu

Bộ dữ liệu **finaleGYS.xlsx** gồm 930 dòng 24 cột, được mô tả cụ thể như bảng sau:

Tên thuộc tính	Mô tả	Kiểu dữ liệu
Rank	Thứ hạng của các kênh trên Youtube	int64
Youtuber	Tên của kênh Youtube	object
Subscribers	Số lượng người đăng ký kênh	int64

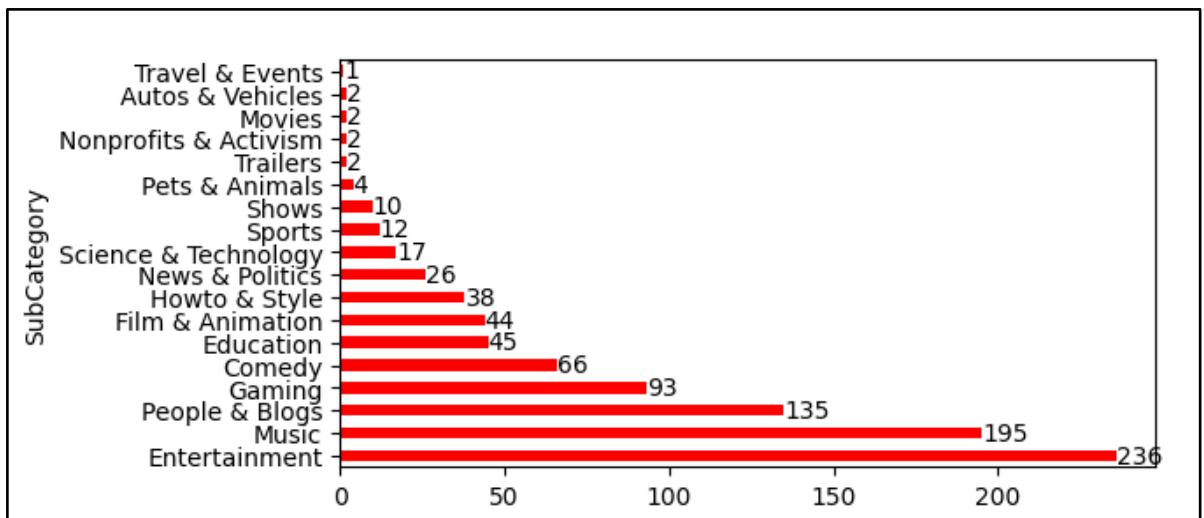
Video Views	Tổng số lượt xem trên tất cả các video của kênh	int64
Uploads	Tổng số lượng video được đăng tải trên kênh	int64
Category	Danh mục của kênh	object
SubCategory	Các thể loại khác của danh mục	object
Country	Quốc gia mà kênh được tạo	object
Region	Châu lục của quốc gia mà kênh được tạo	object
TypeofYoutuber	Hình thức hoạt động của kênh	object
Video Views Rank	Xếp hạng kênh dựa vào tổng lượt xem	int64
Country Rank	Xếp hạng kênh dựa vào số lượng người đăng ký trong nước	int64
Channel Type Rank	Xếp hạng kênh dựa vào chủ đề của kênh	int64
Lowest Monthly Earnings	Ước tính thu nhập hàng tháng thấp nhất từ kênh	float64

Highest Monthly Earning	Ước tính thu nhập hàng tháng cao nhất từ kênh	float64
Lowest Yearly Earnings	Ước tính thu nhập hàng năm thấp nhất từ kênh	float64
Highest Yearly Earning	Ước tính thu nhập hàng năm cao nhất từ kênh	float64
Created Year	Năm mà kênh được tạo	int64
Created Month	Tháng mà kênh được tạo	object
Created Date	Ngày mà kênh được tạo	int64
Gross tertiary education enrollment (%)	Tỷ lệ dân số theo học đại học trong nước	float64
Unemployment rate	Tỷ lệ thất nghiệp của quốc gia	object
Population	Tổng dân số của quốc gia	int64
Urban_population	Tỷ lệ dân số sống ở khu vực đô thị	int64

3.2.4. Thống kê mô tả sau tiền xử lý

3.2.4.1. Mô tả các biến định tính

3.2.4.1.1. Subcategory

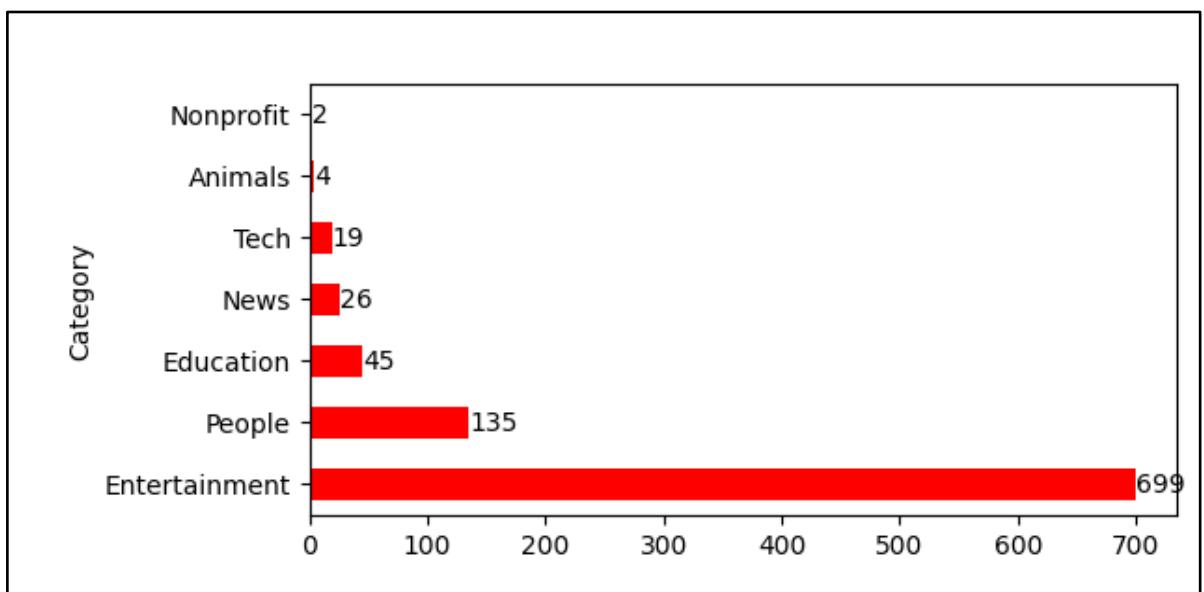


Hình 66 - Biểu đồ thể hiện tần số của các giá trị trong biến Subcategory

Nhân xét:

- Ở Subcategory thì Entertainment là chủ đề có số lượng nhiều nhất (236 kênh), tiếp theo là Music (195 kênh), People & Blogs (135 kênh), Gaming (93 kênh) và Comedy (66 kênh).
- Những danh mục phụ này đều có số lượng lớn cho thấy thị trường YouTube khá ưa chuộng các nội dung giải trí. Mặc dù có sự chênh lệch nhiều, biểu đồ cũng cho thấy sự đa dạng của nội dung trên YouTube, từ giải trí đến giáo dục, khoa học, và nhiều lĩnh vực khác.

3.2.4.1.2. Category

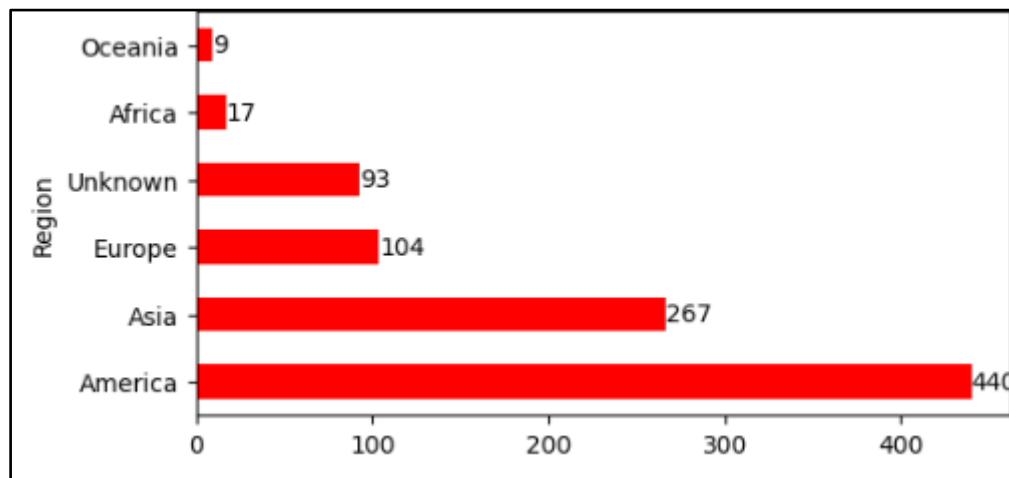


Hình 67 - Biểu đồ thể hiện tần số của các giá trị trong cột Category

Nhân xét:

- Danh mục chiếm số lượng nhiều nhất là Entertainment với 699 kênh, tiếp theo lần lượt là People (135), Education (45), News (26), Tech (19).
- Và thấp nhất là hai kênh có số lượng ít ỏi là Animals với 4 kênh và Nonprofit 2 kênh.

❖ Region

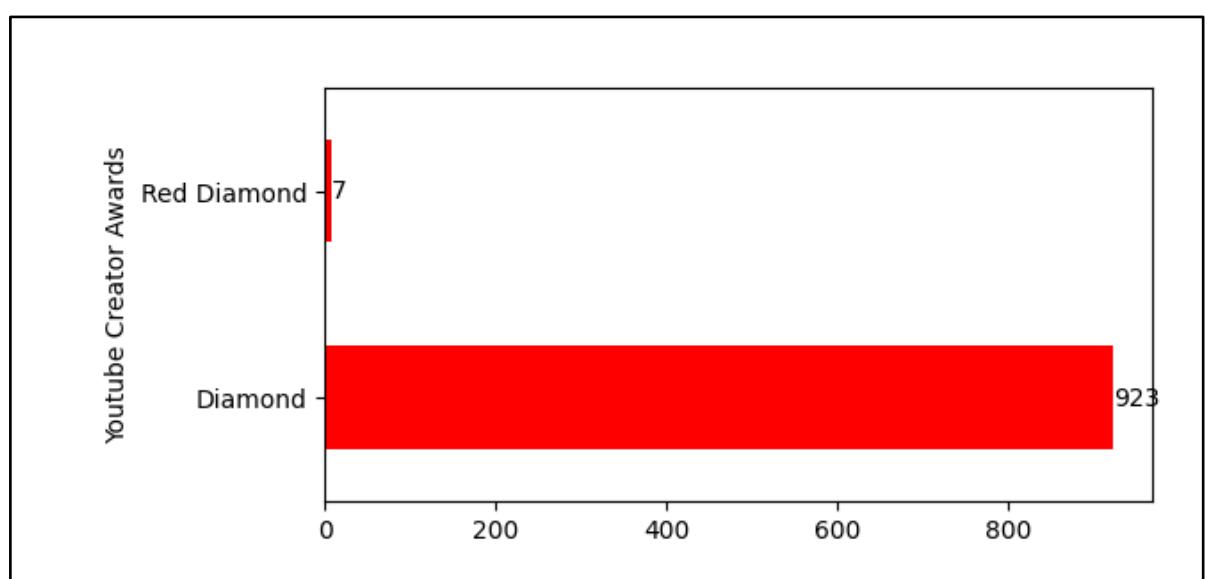


Hình 68 - Biểu đồ thể hiện tần số của các giá trị trong cột Region

Nhân xét:

- Châu Mỹ là khu vực có số lượng nhà sáng tạo nội dung (creator) trên Youtube cao nhất với 440 creators, con số này gấp đôi Châu Á trong khi họ chỉ có 267 creators.
- Số lượng Youtubers thấp gấp 4 lần Châu Mỹ, và gấp hơn 2 lần Châu Á với 104 creators. Ngoài những Youtubers chưa xác định được vùng thì xếp hạng thấp nhất lần lượt là Africa với 17 creators và Oceania có 9 creators.

3.2.4.1.3. Youtube Creator Awards

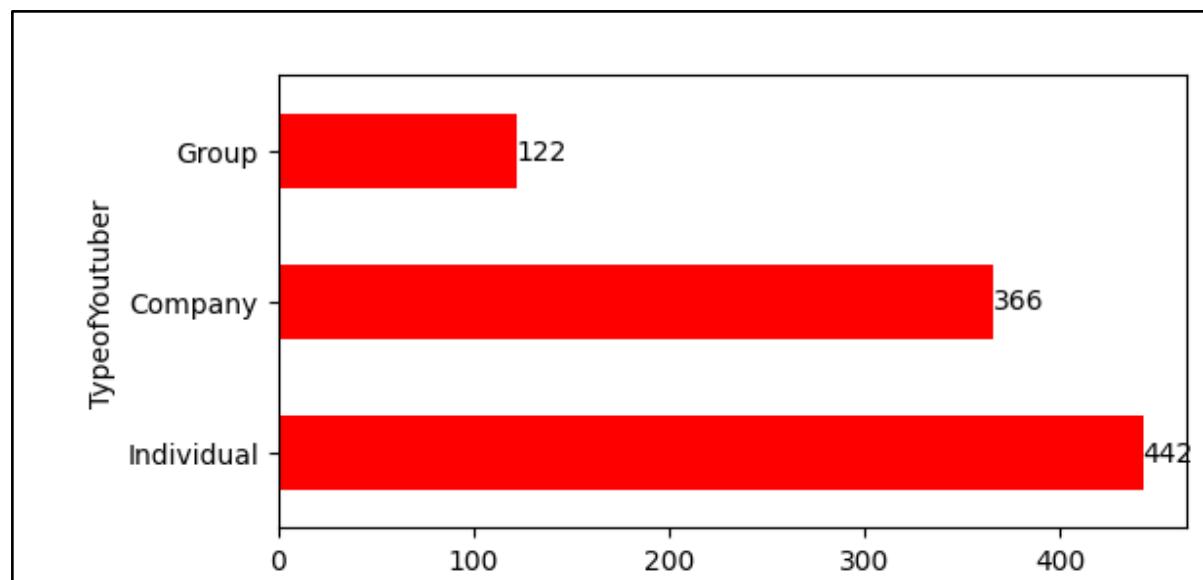


Hình 69 - Biểu đồ thể hiện tần số của các giá trị trong cột Youtube Creator Awards

Nhận xét:

- Nút đỏ Red Diamond là giải thưởng cao quý nhất của Youtube dành tặng cho những youtubers có hơn 100 triệu người theo dõi. Chính vì vậy số lượng của nó chỉ dừng lại ở 7 creators.
- Phần đông trong dataset này được trao tặng nút Diamond, có tận 923 creators đã nhận được nó.

3.2.4.1.4. Type of Youtuber



Hình 70 - Biểu đồ thể hiện tần số của các giá trị trong cột TypeofYoutuber

Nhận xét:

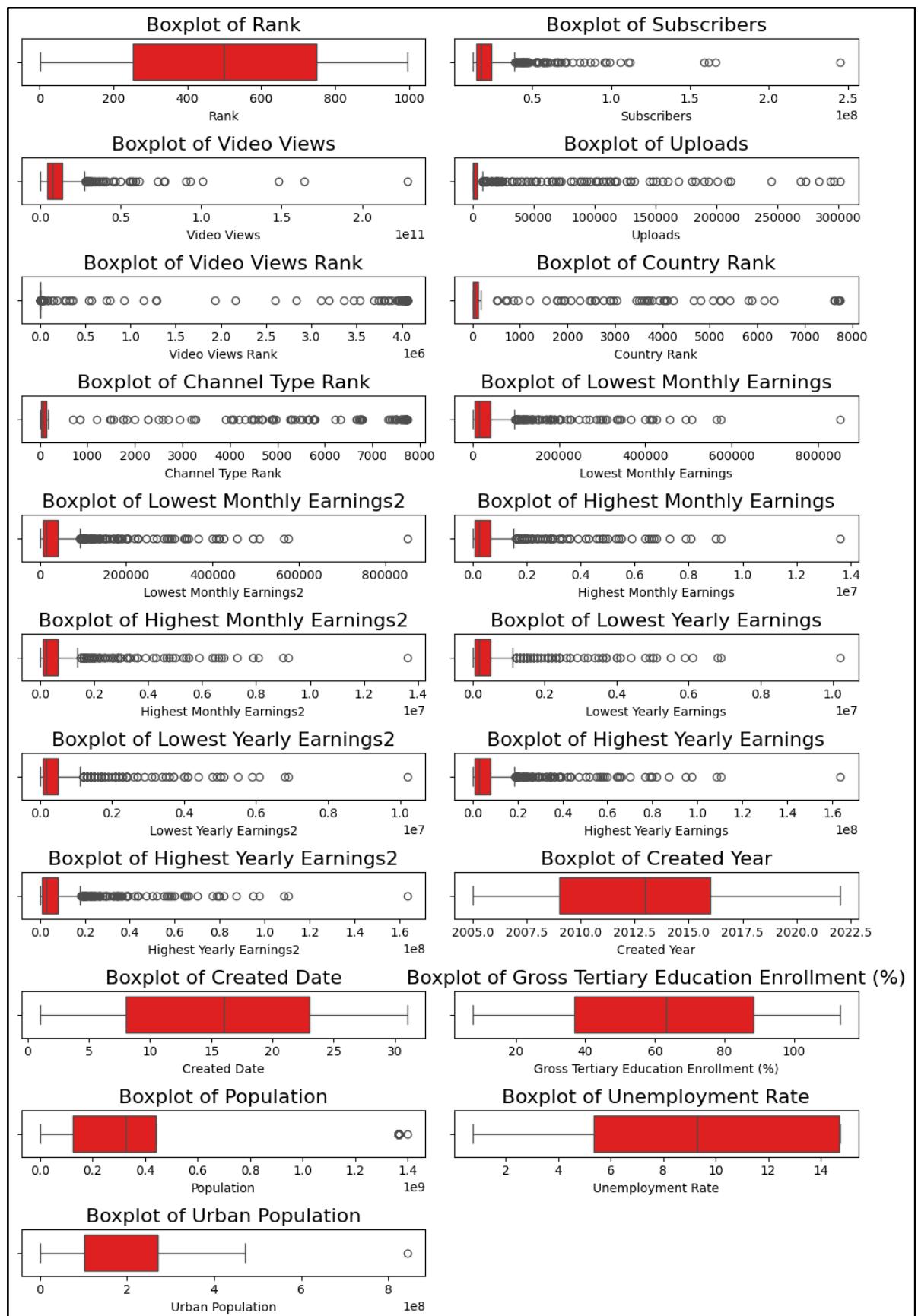
- Youtuber làm cá nhân và Youtuber làm theo công ty có số lượng khá tương đương nhau, tuy nhiên Youtuber cá nhân nhiều hơn 76 kênh so với Youtuber công ty.
- Còn Youtuber làm theo nhóm ít hơn hẳn, chỉ có khoảng 122 kênh, thấp hơn gấp 3.5 lần Youtuber cá nhân.

3.2.4.2. Mô tả các biến định lượng

Phân tích các đại lượng thống kê mô tả là một bước quan trọng trong việc khai thác thông tin từ các biến của bộ dữ liệu.

Nội dung của chương này xoay quanh độ phân tán của các biến cùng với các đại lượng mô tả xu thế trung tâm như trung bình, trung vị, và yếu vị.

3.2.4.2.1. Độ phân tán



Hình 71 - Biểu đồ boxplot thể hiện độ phân tán của các biến numeric

Quan sát các boxplot, những biến có hình dạng phân phối đối xứng gồm: Rank, Created Year, Created Date, Gross Tertiary Education Enrollment. Hình dáng phân phối của các biến numeric còn lại đa phần đều lệch trái, cụ thể:

- Lệch trái gồm các biến: 'Subscribers', 'Video Views', 'Uploads', 'Video Views Rank', 'Country Rank', 'Channel Type Rank', 'Lowest Monthly Earnings', 'Highest Monthly Earnings', 'Lowest Yearly Earnings', 'Highest Yearly Earnings', 'Population', 'Urban Population'. có nhiều dữ liệu rơi vào phía phải của trực, nhưng có một số giá trị “cực thấp” kéo dài về phía trái. Trong trường hợp này, giá trị trung bình và trung vị thường nhỏ hơn mode.
- Lệch phải gồm biến: 'Unemployment Rate'. Có nghĩa là có nhiều dữ liệu rơi vào phía trái của trực, nhưng có một số giá trị “cực cao” kéo dài về phía phải. Trong trường hợp này, giá trị trung bình và trung vị thường lớn hơn mode.

Đồng thời trong số này, một số biến có xuất dữ liệu nhiễu (outliers): Subscribers, Video Views, Uploads, Video View Rank, Country Rank, Channel Type Rank, Lowest Monthly Earnings, Highest Monthly Earnings, Lowest Yearly Earnings, Highest Yearly Earnings, Population, Urban Population.

3.2.4.2.2. Xu thế trung tâm

3.2.4.2.2.1. Rank

Các đại lượng về xu thế trung tâm của biến Rank
Mean: 499.632
Mode: [None]
Median: 499.5

Biến Rank chứa thông tin xếp hạng của các kênh, với khoảng giá trị theo thứ tự từ 1 đến 930. Vì vậy các đại lượng xu thế trung tâm như trung bình, trung vị hay yếu vị không có ý nghĩa đại diện cho đặc trưng của biến này. Cũng có nghĩa rằng phân tích các đại lượng này của Rank là không cần thiết.

3.2.4.2.2.2. Subscribers

Các đại lượng về xu thế trung tâm của biến Subscribers
Mean: 22749677.419

Mode: [12500000.0]

Median: 17700000.0

Trung bình số lượng người đăng ký của một kênh trong bộ dữ liệu này là 22,7 triệu lượt đăng ký. Yếu vị của biến này là 12,500,000. Điều này có nghĩa là những kênh sở hữu 12.5 triệu người đăng ký chiếm ưu thế về số lượng. Trung vị đạt giá trị 17,7 triệu. Tức 50% số kênh có số lượng người đăng ký nhỏ hơn hoặc bằng 17.7 triệu

3.2.4.2.2.3. Video Views

Các đại lượng về xu thế trung tâm của biến Video Views

Mean: 11286933933.213

Mode: [None]

Median: 7804772507.5

Các đại lượng về độ phân tán biến Video Views

Khoảng biến thiên (Range): 227999997366.0

Phương sai (Variance): 2.0882960081846056e+20

Độ lệch chuẩn (Standard deviation): 14450937714.158

Trung bình lượt xem của các kênh rơi vào khoảng hơn 11 tỷ view. Video Views không có giá trị mode bởi bộ dữ liệu có 930 kênh với 930 giá trị lượt xem khác nhau. Trung vị của biến này xấp xỉ 7.804.772.507, nghĩa là một nửa số kênh có lượt xem video nhỏ hơn hoặc bằng 7.8 tỷ

3.2.4.2.2.4. Video Views

Các đại lượng về xu thế trung tâm của biến Uploads

Mean: 10337.739

Mode: [287.0]

Median: 956.0

Các đại lượng về độ phân tán biến Uploads

Khoảng biến thiên (Range): 301297.0

Phương sai (Variance): 1280608081.442

Độ lệch chuẩn (Standard deviation): 35785.585

Trung bình số lượng video mà những kênh trong bộ dữ liệu đã tải lên là khoảng 10 nghìn video. Trong khi đó, giá trị xuất hiện nhiều nhất trong dữ liệu là 287, nghĩa là những kênh đã đăng tải 287 video xuất hiện nhiều nhất. Median của biến Uploads là 956. Điều này có nghĩa là 50% số kênh có số lượng video tải lên nhỏ hơn hoặc bằng 956.

Sự chênh lệch đáng kể của trung bình và trung vị của biến này chỉ ra rằng dữ liệu có một số giá trị rất lớn (các kênh với số lượng video tải lên rất lớn) làm tăng giá trị trung bình. Những giá trị này được gọi là các ngoại lệ và chúng có thể làm “lệch” phân phối của dữ liệu. Trong trường hợp này, dữ liệu có thể bị lệch phải, với một số lượng nhỏ các kênh có số lượng video tải lên rất lớn.

3.2.4.2.2.5. Video Views Rank

Các đại lượng về xu thế trung tâm của biến Video Views Rank
Mean: 385453.903
Mode: [621.0, 630.0, 2218.0]
Median: 847.5

Các đại lượng về độ phân tán biến Video Views Rank
Khoảng biến thiên (Range): 4057942.0
Phương sai (Variance): 1335132833519.418
Độ lệch chuẩn (Standard deviation): 1155479.482

Trung bình xếp hạng lượt xem video là 385,453.903. Nghĩa là lượt xem của các kênh trong bộ dữ liệu trung bình xếp hạng 385,453. Có ba giá trị xuất hiện nhiều nhất trong dữ liệu về xếp hạng theo lượt xem là 621.0, 630.0, và 2218.0. Trung vị của biến Video Views Rank đạt giá trị xấp xỉ 847, tức là 50% số kênh có xếp hạng lượt xem video thấp hơn hoặc bằng 847.

3.2.4.2.2.6. Country Rank

Các đại lượng về xu thế trung tâm của biến Country Rank
Mean: 266.251
Mode: [0.0]
Median: 35.5

Các đại lượng về độ phân tán biến Country Rank
Khoảng biến thiên (Range): 7741.0
Phương sai (Variance): 969489.589
Độ lệch chuẩn (Standard deviation): 984.627

Trung bình của biến Country Rank là 266.251. Điều này nghĩa là các kênh này trung bình xếp hạng 266 trong quốc gia của họ, dựa trên lượt xem. Median của biến này xếp hạng 35, có nghĩa là 50% số kênh xếp hạng thấp hơn hoặc bằng hạng 35 trong xếp hạng quốc gia.

3.2.4.2.2.7. Channel Type Rank

Các đại lượng về xu thế trung tâm của biến Channel Type Rank
Mean: 615.763
Mode: [1.0, 3.0, 8.0, 14.0, 36.0]
Median: 62.0

Các đại lượng về độ phân tán biến Channel Type Rank
Khoảng biến thiên (Range): 7733.0
Phương sai (Variance): 3073381.804
Độ lệch chuẩn (Standard deviation): 1753.106

Trung bình xếp hạng theo loại kênh là 615.763. Đây là xếp hạng trung bình của tất cả các kênh trong dữ liệu. Có năm giá trị xếp hạng xuất hiện nhiều nhất trong dữ liệu là 1, 3, 8, 14 và 36. Trung vị, là 62. Điều này có nghĩa là 50% số kênh có xếp hạng loại kênh nhỏ hơn hoặc bằng 62.

3.2.4.2.2.8. Lowest Monthly Earnings

Các đại lượng về xu thế trung tâm của biến Lowest Monthly Earnings
Mean: 40260.644
Mode: [14700.0]
Median: 14700.0

Các đại lượng về độ phân tán biến Lowest Monthly Earnings
Khoảng biến thiên (Range): 850899.99
Phương sai (Variance): 5363594021.318

Độ lệch chuẩn (Standard deviation): 73236.562

Trung bình thu nhập hàng tháng thấp nhất là 40,260.644. Đây là trung bình thu nhập hàng tháng của tất cả các kênh trong bộ dữ liệu. Giá trị xuất hiện nhiều nhất trong dữ liệu là 14,700. Điều này có nghĩa là kênh có 14,700 đô thu nhập hàng tháng thấp nhất có tần suất xuất hiện nhiều nhất trong bộ dữ liệu. Trung vị, trùng với mode là 14,700. Điều này có nghĩa là 50% số kênh có thu nhập hàng tháng thấp nhất nhỏ hơn hoặc bằng 14,700\$.

Có thể thấy sự phân tán, chênh lệch lớn trong thu nhập hàng tháng thấp nhất giữa các kênh khác nhau.

3.2.4.2.2.9. Highest Monthly Earnings

Các đại lượng về xu thế trung tâm của biến Highest Monthly Earnings

Mean: 636975.05

Mode: [235450.0]

Median: 235475.0

Các đại lượng về độ phân tán biến Highest Monthly Earnings

Khoảng biến thiên (Range): 13599999.99

Phương sai (Variance): 1377529029787.074

Độ lệch chuẩn (Standard deviation): 1173681.826

Mean của biến Highest Monthly Earnings là gần 637 đô. Vậy, các kênh được thống kê có thu nhập cao nhất trung bình đạt hơn 600\$/tháng. Yếu vị và trung vị đều là 235,450\$, tức là những kênh có thu nhập hàng tháng cao nhất đạt 235,450\$ xuất hiện nhiều hơn cả. Đồng thời một nửa số kênh kiểm được nhiều hơn hoặc bằng con số 235\$/tháng.

3.2.4.2.2.10. Lowest Yearly Earnings

Các đại lượng về xu thế trung tâm của biến Lowest Yearly Earnings

Mean: 477611.744

Mode: [176550.0]

Median: 176575.0

Các đại lượng về độ phân tán biến Lowest Yearly Earnings

Khoảng biến thiên (Range): 10199999.99

Phương sai (Variance): 774396155325.297

Độ lệch chuẩn (Standard deviation): 879997.816

Trung bình thu nhập hàng năm thấp nhất là 477,611.744. Giá trị xuất hiện nhiều nhất trong dữ liệu là 176,550. Điều này có nghĩa là có nhiều kênh nhất có thu nhập hàng năm thấp nhất là 176,550. Median tiếp tục trùng với Mode, vậy 50% số kênh có thu nhập hàng năm thấp nhất nhỏ hơn hoặc bằng 176,575\$.

3.2.4.2.2.11. Highest Yearly Earnings

Các đại lượng về xu thế trung tâm của biến Highest Yearly Earnings

Mean: 7619505.487

Mode: [2800000.0]

Median: 2800000.0

Các đại lượng về độ phân tán biến Highest Yearly Earnings

Khoảng biến thiên (Range): 163399999.95

Phương sai (Variance): 199107468531245.7

Độ lệch chuẩn (Standard deviation): 14110544.587

Các kênh có trung bình thu nhập hàng năm cao nhất đạt hơn 7,6 triệu đô. Trong khi đó, thu nhập xuất hiện nhiều nhất là 2,800,000\$/ năm. Đây cũng là giá trị trung vị của biến thu nhập hàng năm cao nhất.

3.2.4.2.2.12. Created Year

Các đại lượng về xu thế trung tâm của biến Created Year

Mean: 2012.683

Mode: [2014.0]

Median: 2013.0

Các đại lượng về độ phân tán biến Created Year

Khoảng biến thiên (Range): 17.0

Phương sai (Variance): 18.344

Độ lệch chuẩn (Standard deviation): 4.283

Xu hướng trung tâm của biến Created Year có giá trị trung bình là 2012, giá trị trung vị là 2013, và mode là 2014. Điều này cho thấy rằng phần lớn các kênh YouTube trong dữ liệu của được tạo vào khoảng năm 2012 đến 2014.

3.2.4.2.2.13. Created Date

Các đại lượng về xu thế trung tâm của biến Created Date

Mean: 15.775

Mode: [9.0, 19.0]

Median: 16.0

Các đại lượng về độ phân tán biến Created Date

Khoảng biến thiên (Range): 30.0

Phương sai (Variance): 76.971

Độ lệch chuẩn (Standard deviation): 8.773

Xu hướng trung tâm của biến Created Date có giá trị trung bình là gần 16, giá trị trung vị là 16, và mode gồm 9 và 19. Điều này cho thấy rằng các kênh YouTube trong dữ liệu được tạo vào nhiều nhất vào ngày 9, 19.

3.2.4.2.2.14. Gross Tertiary Education

Các đại lượng về xu thế trung tâm của biến Gross Tertiary Education

Enrollment (%)

Mean: 63.109

Mode: [88.2]

Median: 63.1

Các đại lượng về độ phân tán biến Gross Tertiary Education

Enrollment (%)

Khoảng biến thiên (Range): 105.5

Phương sai (Variance): 616.279

Độ lệch chuẩn (Standard deviation): 24.825

Tỷ lệ theo học đại học trung bình của các quốc gia xuất hiện trong bộ dữ liệu xấp xỉ 63%. Quốc gia có 88.2% có tần suất xuất hiện nhiều nhất.

3.2.4.2.2.15. Population

Các đại lượng về xu thế trung tâm của biến Population

Mean: 440330922.152

Mode: [328239523.0]

Median: 328239523.0

Các đại lượng về độ phân tán biến Population

Khoảng biến thiên (Range): 1397512494.0

Phương sai (Variance): 2.053236285113284e+17

Độ lệch chuẩn (Standard deviation): 453126503.872

Trung bình dân số dựa trên các quốc gia của bộ dữ liệu là 440,330,922. Trong khi đó trung vị tính được là 328,239,523.

3.2.4.2.2.16. Unemployment Rate

Các đại lượng về xu thế trung tâm của biến Unemployment Rate

Mean: 9.302

Mode: [14.7]

Median: 9.3

Các đại lượng về độ phân tán biến Unemployment Rate

Khoảng biến thiên (Range): 13.97

Phương sai (Variance): 21.558

Độ lệch chuẩn (Standard deviation): 4.643

Tỷ lệ thất nghiệp của các quốc gia trong bộ dữ liệu là 9,3%. Trong đó, giá trị xuất hiện nhiều nhất là 14,7%

3.2.4.2.2.17. Urban Population

Các đại lượng về xu thế trung tâm của biến Urban Population

Mean: 227682635.955

Mode: [270663028.0]

Median: 270663028.0

Các đại lượng về độ phân tán biến Urban Population

Khoảng biến thiên (Range): 842898374.0

Phương sai (Variance): 2.1637989027044628e+16

Độ lệch chuẩn (Standard deviation): 147098569.086

Dân số thành thị trung bình của các quốc gia trong bộ dữ liệu là khoảng 227,682,635. Median đạt 270,663,028.

Nhận xét: Đa phần những biến về hiệu suất kinh, xếp hạng và thu nhập có mean và median chênh lệch nhiều, có độ biến thiên và mức độ phân tán dữ liệu lớn, xuất hiện nhiều dữ liệu nhiễu. Trong trường hợp này, median có thể là một đại diện tốt hơn cho xu thế trung tâm của dữ liệu, vì nó không bị ảnh hưởng bởi các giá trị ngoại lệ.

3.2.5. Tạo các bảng dimension, fact

3.2.5.1. dimTime

Tên thuộc tính	Mô tả
TimeID	<p>Khóa chính của bảng dimTime, có 6 ký tự theo dạng “DDMMYY”, được tạo dựa theo cột Time.</p> <p>Công thức tạo cột TimeID trong Excel:</p> $=TEXT(B2,"DDMMYY")$ <p>Trong đó B2 là cột Time.</p>
Time	Các giá trị được gom lại từ 3 cột Created Date, Created Month và Created Year của bảng OriginalGYS, sau đó loại bỏ dòng trùng để lấy giá trị duy nhất.
MonthID	Lấy 2 chữ số của tháng (ví dụ: 01, 06, 12).
YearID	Lấy 2 chữ số của năm (ví dụ: 2005 sẽ có YearID là 05).
DateID	Lấy 2 chữ số của ngày (ví dụ: 01, 19, 29).

Bảng 5 - Mô tả cách tạo các cột trong bảng dimTime

3.2.5.2. dimYear

Tên thuộc tính	Mô tả

YearID	<p>Khóa chính của bảng dimYear, có 2 ký tự.</p> <p>Công thức tạo cột YearID trong Excel:</p> <p style="padding-left: 40px;"><code>=RIGHT(B2,2)</code></p> <p>Trong đó B2 là cột CreatedYear</p>
CreatedYear	Các giá trị được lấy từ cột Created Year của bảng OriginalGYS, sau đó bỏ dòng trùng để lấy giá trị duy nhất.

Bảng 6 - Mô tả cách tạo các cột trong bảng dimYear

3.2.5.3. dimMonth

Tên thuộc tính	Mô tả
MonthID	Khóa chính của bảng dimMonth, có 2 ký tự. Do cột Month có dạng Tên số lượng tháng không nhiều là 12. Vì thế nhóm tạo khóa chính bằng cách công.
CreatedMonth	Các giá trị được lấy từ cột Created Month của bảng OriginalGYS, sau đó bỏ dòng trùng để lấy giá trị duy nhất.

Bảng 7 - Mô tả cách tạo các cột trong bảng dimMonth

3.2.5.4. dimRegion

Tên thuộc tính	Mô tả
RegionID	<p>Khóa chính của bảng dimRegion. Khóa chính được tạo bằng cách lấy đầu trong cột Region.</p> <p>Công thức tạo cột RegionID trong Excel:</p> <p style="padding-left: 40px;"><code>=UPPER(LEFT(B2,2))</code></p> <p>Trong đó B2 là cột Region</p>
Region	Các giá trị được lấy từ cột Region của bảng OriginalGYS, sau đó loại bỏ trùng để lấy giá trị duy nhất.

Bảng 8 - Mô tả cách tạo các cột trong bảng dimRegion

3.2.5.5. dimCountry

Tên thuộc tính	Mô tả
CountryID	Khóa chính của dimCountry. Khóa chính gồm 2 ký tự, là viết tắt của các giá trị trong cột Country theo tiêu chuẩn ISO 3166-1 alpha-2.
Country	Các giá trị được lấy từ cột Country của bảng OriginalGYS, sau đó loại bỏ dòng trùng để lấy giá trị duy nhất.
RegionID	Lấy 2 ký tự đầu của khu vực đó.

Bảng 9 - Mô tả cách tạo các cột trong bảng dimCountry

3.2.5.6. dimCategory

Tên thuộc tính	Mô tả
CategoryID	<p>Khóa chính của bảng dimCategory. Khóa chính được tạo bằng cách lấy đầu trong cột Category.</p> <p>Công thức tạo cột CategoryID trong Excel:</p> $=UPPER(LEFT(B2,2))$ <p>Trong đó B2 là cột Category</p>
Category	Các giá trị được lấy từ cột Category của bảng OriginalGYS, sau đó loại bỏ trùng để lấy giá trị duy nhất.

Bảng 10 - Mô tả cách tạo các cột trong bảng dimCategory

3.2.5.7. dimSubcategory

Tên thuộc tính	Mô tả
SubcategoryID	<p>Khóa chính của bảng dimSubcategory.</p> <ul style="list-style-type: none"> • Nếu Subcategory chỉ có 1 từ (ví dụ: Music), khóa chính là 3 ký tự đầu • Nếu Subcategory có 2 từ (ví dụ: People & Blogs), khóa chính là 4 ký tự đầu

	<p>chính có dạng <ký tự đầu của từ thứ nhất + A + ký tự đầu của từ thứ hai></p> <p>Công thức tạo cột SubcategoryID trong Excel:</p> <pre>=UPPER(IF(ISNUMBER(SEARCH("&",B2)),CONCATENATE(LEFT(B2,1),"A",LEFT(TEXTAFTER(B2,"& ",1,1,0,"none"),1)))</pre> <p>Trong đó B2 là cột Subcategory</p>
Subcategory	Các giá trị được lấy từ cột Category của bảng OriginalGYS, sau đó loại bỏ dòng trùng để lấy giá trị duy nhất.
CategoryID	Lấy 2 ký tự đầu của Category

Bảng 11 - Mô tả cách tạo các cột trong bảng dimSubcategory

3.2.5.8. dimTypeofYoutuber

Tên thuộc tính	Mô tả
TypeID	<p>Khóa chính của bảng dimTypeOfYoutuber. Khóa chính được tạo bằng 3 ký tự đầu trong cột TypeOfYoutuber.</p> <p>Công thức tạo cộtTypeID trong Excel:</p> <pre>=UPPER(LEFT(B2,3))</pre> <p>Trong đó B2 là cột TypeOfYoutuber</p>
TypeofYoutuber	Các giá trị được lấy từ cột TypeOfYoutuber của bảng OriginalGYS, bỏ dòng trùng để lấy giá trị duy nhất.

Bảng 12 - Mô tả cách tạo các cột trong bảng dimTypeofYoutuber

3.2.5.9. dimYoutuberCreatorAwards

Tên thuộc tính	Mô tả
AwardsID	<p>Khóa chính của bảng dimYoutuberCreatorAwards.</p> <ul style="list-style-type: none"> Nếu YoutuberCreatorAwards chỉ có 1 từ (ví dụ: Diamond), khóa chính kí tự đầu Nếu YoutuberCreatorAwards có 2 từ (ví dụ: Red Diamond, khóa chín

	<p>dạng <ký tự đầu của từ thứ nhất + ký tự đầu của từ thứ hai></p> <p>Công thức tạo cột AwardsID trong Excel:</p> <pre>=UPPER(IF(ISNUMBER(SEARCH(" ",B2)),CONCATENATE(LEFT(B2,1),LEFT(TEXTAFTER(B2, " ",1,1,0,"none"),1)),LEFT(B2,2)))</pre> <p>Trong đó B2 là cột YoutuberCreatorAwards</p>
YoutubeCreatorAwards	Các giá trị được lấy từ cột YoutuberCreatorAwards bảng OriginalGYS đó loại bỏ dòng trùng để lấy giá trị duy nhất.

Bảng 13 - Mô tả cách tạo các cột trong bảng dimYoutuberCreatorAwards

3.2.5.10. factGYS

Đầu tiên, dữ liệu được sao chép từ bảng OriginalGYS đến bảng FactGYS.

Các cột bị loại bỏ **Created Date, Created Month, Created Year, Region, Country, Category, Subcategory, TypeOfYoutuber** và **YoutuberCreatorAwards**.

Các cột thêm vào:

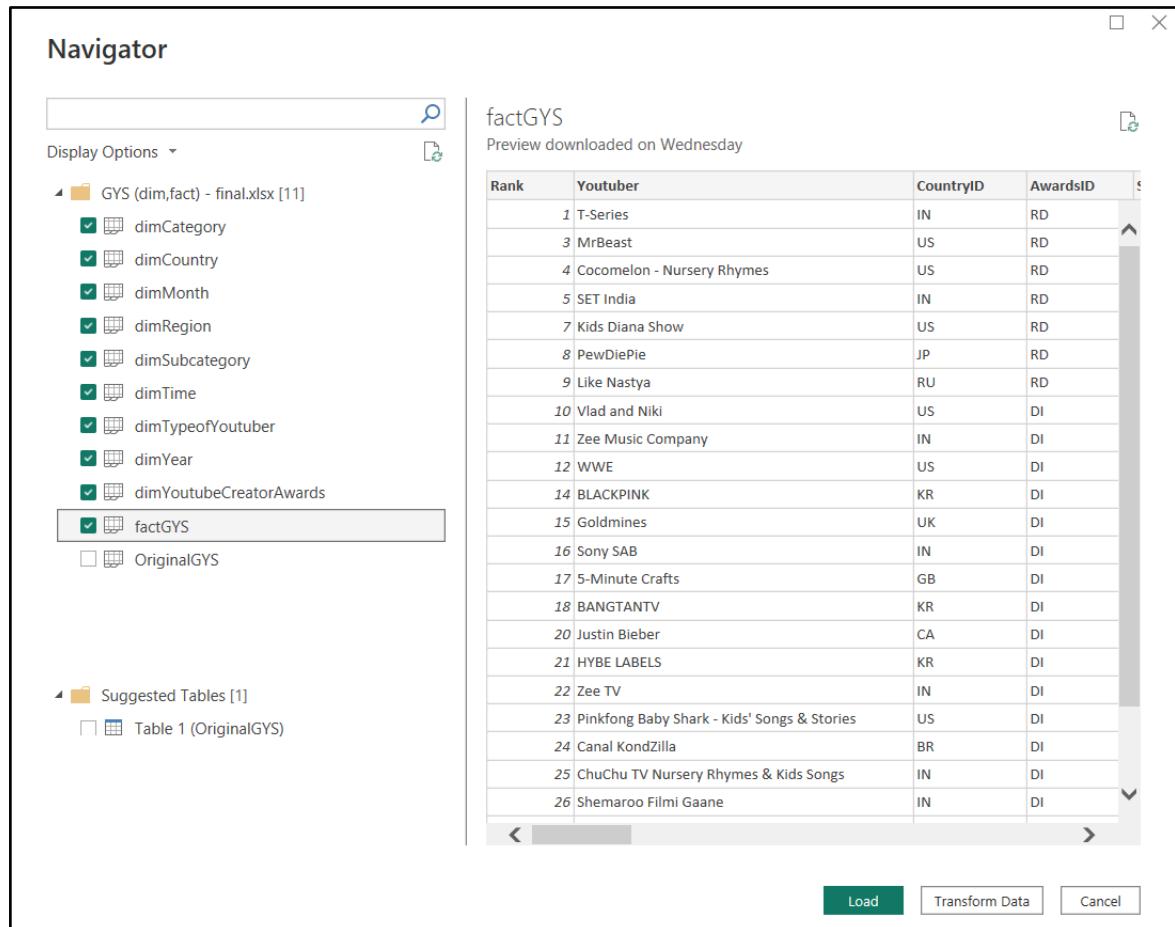
Thuộc tính	Mô tả
TimeID	LOOKUP để lấy các giá trị từ cột TimeID của bảng dimTime theo lần lượt các giá trị trong cột Time (được tạo từ Created Date, Created Month, Created Year) của bảng OriginalGYS
CountryID	LOOKUP để lấy các giá trị từ cột CountryID của bảng dimCountry theo lần lượt các giá trị trong cột Country bảng OriginalGYS
SubcategoryID	LOOKUP để lấy các giá trị từ cột SubcategoryID của bảng dimSubcategory theo lần lượt các giá trị trong cột Subcategory bảng OriginalGYS
TypeID	LOOKUP để lấy các giá trị từ cộtTypeID của bảng dimTypeOfYoutuber theo lần lượt các giá trị trong cột TypeOfYoutuber bảng OriginalGYS
AwardsID	LOOKUP để lấy các giá trị từ cột AwardsID của bảng dimYoutuberCreatorAwards theo lần lượt các giá trị trong cột YoutuberCreatorAwards bảng OriginalGYS

Bảng 14 - Mô tả cách tạo các cột trong bảng factGYS

Các cột có biến measurable và cột Youtuber được giữ nguyên.

3.2.6. Chuyển đổi dữ liệu trong Power BI

Để Load dữ liệu vào Power BI, chọn **Import data from Excel** và chọn các bảng cần thiết. Tuy nhiên, trước khi **Load** dữ liệu cần kiểm tra và điều chỉnh các thông tin của các bảng như tiêu đề cột, kiểu dữ liệu nên cần chọn **Transform Data**.



Hình 72 - Nhập dữ liệu trong file Excel vào PowerBI

3.2.6.1. Chính sửa tiêu đề cột

Lần lượt kiểm tra từng bảng, nhóm nhận thấy một số bảng đang có tiêu đề cột không đúng, bao gồm các bảng **dimCategory**, **dimCountry**, **dimRegion**, **dimSubcategory**, **dimTypeofYoutuber** và **dimYoutubeCreatorAwards**. Tiêu đề cột được đặt là Column 1, Column 2... trong khi tên cột chính xác lại nằm ở dòng 1.

Ví dụ ở bảng **dimCategory**:

Queries [10]

dimCategory

CategoryID	Category
EN	Entertainment
ED	Education
PE	People
NE	News
NO	Nonprofit
TE	Tech
AN	Animals

Hình 73 - Bảng dimCategory trong PowerBI

Để chỉnh sửa, nhấp chuột phải và chọn **Use First Row as Headers**.

Queries [10]

dimCategory

CategoryID	Category
EN	Entertainment
ED	Education
PE	People
NE	News
NO	Nonprofit
TE	Tech
AN	Animals

Hình 74 - Xử lý bảng dimCategory trong PowerBI

Kết quả:

Queries [10]

dimCategory

CategoryID	Category
EN	Entertainment
ED	Education
PE	People
NE	News
NO	Nonprofit
TE	Tech
AN	Animals

Hình 75 - Kết quả sau khi xử lý dimCategory trong PowerBI

Các bảng còn lại là **dimCountry**, **dimRegion**, **dimSubcategory**, **dimTypeofYoutuber** và **dimYoutubeCreatorAwards** cũng sẽ được chỉnh sửa tương tự.

3.2.6.2. Thay đổi kiểu dữ liệu cột

Đối với các cột ID của các bảng **dimYear** (YearID), **dimMonth** (MonthID), **dimTime** (TimeID) và **factGYS** (TimeID, DateID, MonthID, YearID), hiện có kiểu dữ

liệu là **Whole Number** và được định dạng như 1, 2...10. Để đồng nhất về chiều dài dữ liệu trong từng cột như 01, 02...10, nhóm tiến hành chuyển đổi kiểu của các cột trên về dạng **Text**.

Ví dụ ở bảng **dimMonth**:

MonthID	CreatedMonth
1	Jan
2	Feb
3	Mar
4	Apr
5	May
6	Jun
7	Jul
8	Aug
9	Sep
10	Oct
11	Nov
12	Dec

Hình 76 - Bảng dimMonth trong PowerBI

Để chuyển đổi kiểu dữ liệu một cột, nhấp chuột phải ở cột đó, chọn **Change Type** và chọn **Text**.

Hình 77 - Xử lý bảng dimMonth trong PowerBI

Sau đó chọn **Replace current**.

Hình 78 - Xử lý bảng dimMonth trong PowerBI

Kết quả:

Hình 79 - Kết quả sau khi xử lý dimMonth trong PowerBI

Các cột còn lại của các bảng **dimYear** (YearID), **dimTime** (TimeID) và **factGYS** (TimeID, DateID, MonthID, YearID) sẽ được thực hiện tương tự. Cuối cùng, nhóm tiến hành **Load** dữ liệu vào Power BI bằng cách chọn **Close & Apply**.

3.3. Giai đoạn 3: Nạp dữ liệu vào kho dữ liệu (Load)

3.3.1. Giới thiệu các bảng Dim

Các dimension: chứa các chi tiết về dữ liệu trong bảng fact, Các bảng này sẽ được nối đến các bảng fact thông qua cột khóa (key column). Bảng Dim chứa các giá trị là duy nhất, chẳng hạn, mỗi dòng trong bảng Country thể hiện một quốc gia duy nhất. Bộ dữ liệu bao gồm 9 bảng dimension và 1 bảng fact.

3.3.1.1. dimTypeofYoutuber

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
TypeID	Mã định danh (ID) hình thức hoạt động của các kênh	text	Các kênh sẽ được phân thành các nhóm khác nhau dựa trên cách thức các kênh hoạt động.
TypeofYoutuber	Tên các hình thức hoạt động	object	Có 3 nhóm lần lượt là: Individual, Group và Company

Bảng 15 - Mô tả các cột trong bảng dimTypeofYoutuber

3.3.1.2. dimCategory

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
CategoryID	Mã danh mục	text	Bao gồm 7 dòng danh mục khác nhau
Category	Tên của từng danh mục	object	Kết nối với dimension "dimSubCategory"

Bảng 16 - Mô tả các cột trong bảng dimCategory

3.3.1.3. dimSubcategory

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
SubcategoryID	Mã danh mục con của kênh	text	
Subcategory	Tên danh mục phụ	object	
CategoryID	Mã danh mục	text	

Bảng 17 - Mô tả các cột trong bảng dimSubcategory

3.3.1.4. dimTime

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú

TimeID	Mã thời gian	text	
Time	Thời gian tạo kênh	int64	
MonthID	Mã của tháng kênh được tạo	text	
YearID	Mã của năm kênh được tạo	text	
DateID	Mã của ngày kênh youtube được tạo	text	

Bảng 18 - Mô tả các cột trong bảng dimTime

3.3.1.5. dimMonth

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
MonthID	Mã các tháng trong năm mà kênh được tạo	text	Kết nối với dimension "dimTime"
Month	Tháng cụ thể mà kênh được tạo	int64	

Bảng 19 - Mô tả các cột trong bảng dimMonth

3.3.1.6. dimYear

Tên thuộc tính	Mô tả	Kiểu	Ghi chú

		dữ liệu	
YearID	Mã năm tạo kênh	text	Kết nối với dimension "dimTime"
Year	Năm cụ thể mà kênh tạo	int64	

Bảng 20 - Mô tả các cột trong bảng dimYear

3.3.1.7. dimCountry

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
CountryID	Gồm 49 dòng dữ liệu chứa mã định danh của các quốc gia.	text	
Country	Tên quốc gia	object	
RegionID	Mã khu vực	text	

Bảng 21 - Mô tả các cột trong bảng dimCountry

3.3.1.8. dimRegion

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú

RegionID	Mã các khu vực của các kênh Youtube	text	Gồm 6 dòng mã khu vực khác nhau Kết nối với dimension "dimCountry" "
Region	Tên khu vực	object	

Bảng 22 - Mô tả các cột trong bảng dimRegion

3.3.1.9. dimYoutubeCreatorAwards

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Ghi chú
AwardsID	Mã giải thưởng mà các kênh đạt được	text	Các giải thưởng được tính trên số lượng subscribers.
YoutubeCreator orAwards	Tên các giải	object	

Bảng 23 - Mô tả các cột trong bảng dimYoutubeCreatorAwards

3.3.2. Giới thiệu bảng Fact

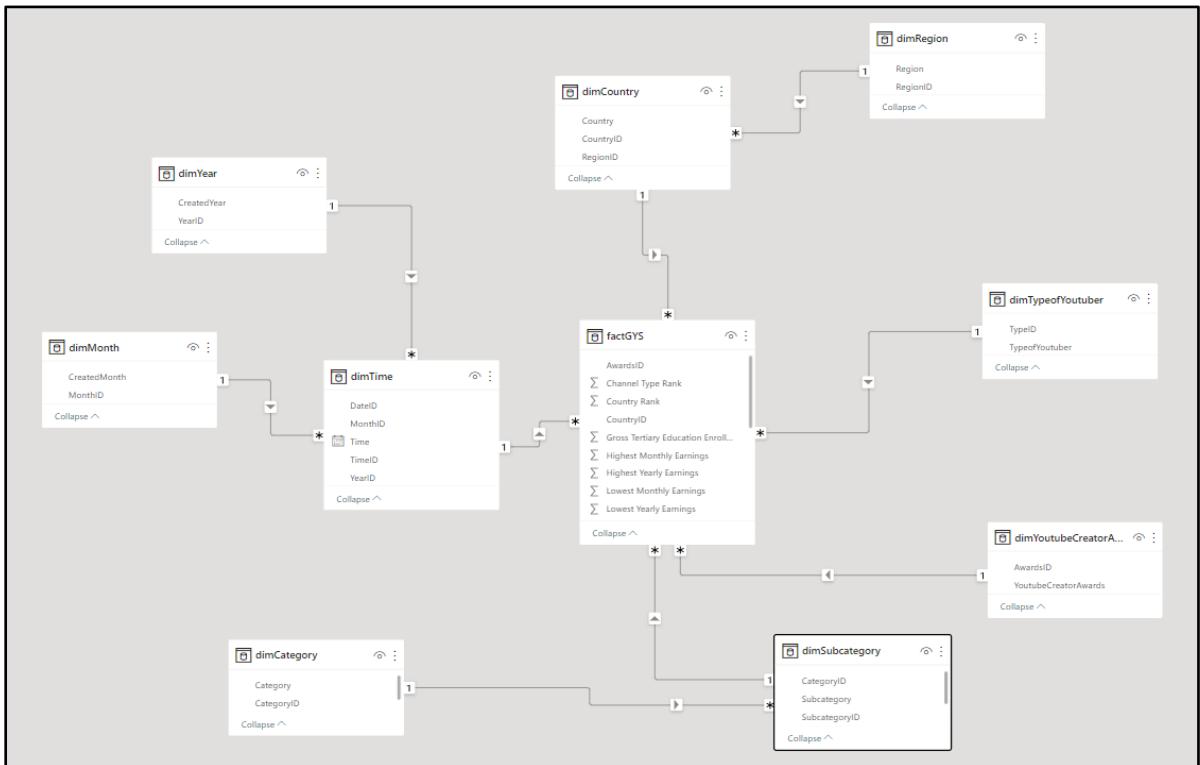
Tên thuộc tính	Mô tả	Loại dữ liệu	Ghi chú
Rank	Thứ hạng của kênh Youtube	int64	
Youtuber	Tên của Youtuber	object	
CountryID	Mã ID quốc gia của kênh được tạo	text	
AwardsID	Mã ID giải thưởng mà kênh đạt được	text	Gồm 2 giá trị: RD và DI, được tính trên số lượng subscribers.
SubcategoryID	Mã ID danh mục con của kênh	text	
TypeID	Mã ID hình thức hoạt động của kênh	text	Được chia làm 3 nhóm: individual, company và group
TimeID	Mã ID thời gian mà kênh được tạo	int64	
Subscribers	Số lượng người đăng ký kênh	int64	

Video Views	Tổng số lượt xem trên tất cả các video của kênh	float64 4	
Uploads	Số lượng video được đăng tải của kênh	int64	
Video Views Rank	Xếp hạng kênh dựa trên tổng số lượt xem video	float64 4	
Country Rank	Xếp hạng kênh dựa trên số lượng người đăng ký trong nước	float64 4	
Channel Type Rank	Xếp hạng kênh dựa vào chủ đề của kênh Youtube	float64 4	
Lowest Monthly Earnings	Ước tính thu nhập hàng tháng thấp nhất của kênh	float64 4	
Highest Monthly Earnings	Ước tính thu nhập hàng tháng cao nhất của kênh	float64 4	
Lowest Yearly	Ước tính thu nhập hàng năm	float64	

Earnings	thấp nhất của kênh	4	
Highest Yearly Earnings	Ước tính thu nhập hàng năm cao nhất của kênh	float6 4	
Gross Tertiary Education Enrollment (%)	Tỷ lệ dân số theo học đại học trong nước	float6 4	
Population	Tổng dân số của quốc gia	float6 4	
Unemployment Rate	Tỷ lệ thất nghiệp trong nước	objec t	
Urban Population	Dân số sống ở thành thị của quốc gia	float6 4	

Bảng 24 - Mô tả các cột trong bảng factGYS

3.3.3. Mô hình dữ liệu (Data Model)



Hình 80 - Mô hình dữ liệu

Mô hình dữ liệu mà nhóm chọn là Lược đồ bông tuyết (Snowflake Schema). Đây là mô hình dữ liệu đa chiều, là mở rộng của lược đồ hình sao. Trong đó, một hoặc nhiều bảng dimension không kết nối trực tiếp với bảng fact mà phải kết nối thông qua các bảng dimension khác. Lược đồ bông tuyết của nhóm bao gồm 9 bảng dimension và 1 bảng fact, trong đó 5 bảng dimension được kết nối trực tiếp với bảng fact bao gồm dimTime, dimCountry, dimTypeofYoutuber, dimYoutubeCreatorAwards và dimSubcategory. Các bảng dimension còn lại được kết nối thông qua 1 bảng dimension khác, cụ thể: dimMonth và dimYear cùng kết nối với dimTime, dimRegion kết nối với dimCountry, dimCategory kết nối với dimSubcategory.

CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU

4.1. Các Measure

4.1.1. dimYear

```
SelectedYear =  
VAR SYear = FORMAT(SELECTEDVALUE(dimYear[CreatedYear]), "0000")  
RETURN  
| IF(ISBLANK(SELECTEDVALUE(dimYear[CreatedYear])), "All Years", SYear)
```

Hình 81 - DAX hiển thị năm được chọn

Measure hiển thị năm được chọn. Nếu không có năm nào được chọn, measure này sẽ trả về giá trị mặc định ‘All year’. Nếu có năm được chọn, nó sẽ hiển thị năm đó.

Thành phần của công thức:

- **Biến SYear:** dùng hàm SELECTEDVALUE để gán năm được chọn, định dạng format theo kiểu ‘yyyy’
- **Hàm SELECTEDVALUE** để lấy năm đã chọn. Nếu không có giá trị nào hoặc nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **Hàm ISBLANK** kiểm tra xem giá trị trả về từ SELECTEDVALUE có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.
- **Điều Kiện IF:** Hàm IF sẽ kiểm tra điều kiện từ ISBLANK. Nếu TRUE, nó sẽ trả về giá trị mặc định ‘All year’. Nếu FALSE, nó sẽ trả về năm đã được chọn.

4.1.2. dimMonth

```
SelectedMonth =  
VAR SMonth = SELECTEDVALUE(dimMonth[CreatedMonth])  
RETURN  
| IF(ISBLANK(SELECTEDVALUE(dimMonth[CreatedMonth])), "All Months", SMonth)
```

Hình 82 - DAX hiển thị tháng được chọn

Measure hiển thị tháng được chọn. Measure hiển thị tháng được chọn. Nếu không có tháng nào được chọn, measure này sẽ trả về giá trị mặc định ‘All Months’. Nếu có tháng được chọn, nó sẽ hiển thị tháng đó.

Thành phần của công thức:

- **Biến SMonth:** dùng hàm SELECTEDVALUE để gán month được chọn

- **Hàm SELECTEDVALUE** để lấy tháng đã chọn. Nếu không có giá trị nào hoặc nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **Hàm ISBLANK** kiểm tra xem giá trị trả về từ SELECTEDVALUE có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.
- **Điều Kiện IF:** Hàm IF sẽ kiểm tra điều kiện từ ISBLANK. Nếu TRUE, nó sẽ trả về giá trị mặc định ‘All Months’. Nếu FALSE, nó sẽ trả về tháng đã được chọn.

4.1.3. dimRegion

- a. Measure đếm số quốc gia có trong khu vực đó

```
1 CountCountriesByRegion = COUNTROWS(SUMMARIZE('factGYS', 'dimCountry'[Country], 'dimRegion'[Region]))
```

Hình 83 - DAX tính số lượng quốc gia

Hàm này **tính số lượng các quốc gia** (countries) dựa trên việc tạo một bảng tóm tắt (**SUMMARIZE**) từ bảng factGYS. Bảng tóm tắt này bao gồm các cột Country và Region từ các bảng liên quan. Sau đó, hàm **COUNTROWS** đếm số dòng trong bảng tóm tắt, tương ứng với số lượng quốc gia.

- b. Measure hiển thị khu vực đang được chọn

```
1 SelectedRegion =
2 VAR SRegion = SELECTEDVALUE(dimRegion[Region])
3 RETURN
4 | IF(ISBLANK(SELECTEDVALUE(dimRegion[Region])), "All Regions", SRegion)
```

Hình 84 - DAX hiển thị khu vực được chọn

- **VAR SRegion = SELECTEDVALUE(dimRegion[Region]):** Dòng này tạo một biến tạm thời có tên là SRegion. Nó gán giá trị của cột Region trong bảng dimRegion khi có lựa chọn (sử dụng SELECTEDVALUE). Nếu không có lựa chọn, biến này sẽ giữ giá trị null.
- **RETURN IF(ISBLANK(SELECTEDVALUE(dimRegion[Region])), "All Regions", SRegion):** Đây là phần trả về của hàm. Nó kiểm tra xem giá trị của biến SRegion có phải là null hay không (sử dụng ISBLANK). Nếu có, hàm sẽ trả về chuỗi “All Regions”. Ngược lại, nó sẽ trả về giá trị của biến SRegion.

4.1.4. dimCountry

- a. Measure hiển thị quốc gia đang được chọn

```

1 SelectedCountry =
2 VAR SCountry = SELECTEDVALUE(dimCountry[Country])
3 RETURN
4 | IF(ISBLANK(SELECTEDVALUE(dimCountry[Country])), "All Countries", SCountry)

```

Hình 85 - DAX hiển thị quốc gia được chọn

- **VAR SCountry = SELECTEDVALUE(dimCountry[Country]):** Đây là một biến tạm thời (VAR) được gán giá trị là giá trị của cột Country trong bảng dimCountry khi có lựa chọn (SELECTEDVALUE). Nếu không có lựa chọn, biến này sẽ giữ giá trị null.
- **RETURN IF(ISBLANK(SELECTEDVALUE(dimCountry[Country])), "All Countries", SCountry):** Đây là phần trả về của hàm. Nó kiểm tra xem giá trị của biến SCountry có null hay không. Nếu có, nó trả về chuỗi “All Countries”, ngược lại nó trả về giá trị của biến SCountry.

b. Measure cho biết Country thuộc Region nào

```

1 CountryofRegion = LOOKUPVALUE(dimRegion[Region], dimRegion[RegionID], SELECTEDVALUE(dimCountry[RegionID]))
2

```

Hình 86 - DAX hiển thị khu vực của Country được chọn

Đây là một hàm dùng để tìm giá trị trong cột Region của bảng dimRegion dựa trên giá trị của cột RegionID trong bảng dimCountry khi có lựa chọn (SELECTEDVALUE). Nếu không có lựa chọn, hàm sẽ trả về giá trị null.

4.1.5. dimCategory

```

1 SelectedCategory = IF(ISBLANK(SELECTEDVALUE(dimCategory[Category])), "All Categories", SELECTEDVALUE(dimCategory[Category]))

```

Hình 87 - DAX hiển thị chủ đề được chọn

Measure hiển thị Category được chọn. Biến **SelectedCategory** sẽ chứa giá trị “All Categories” nếu không có giá trị nào được chọn trong cột Category, hoặc nó sẽ chứa giá trị đã được chọn. Cụ thể các hàm bên trong measure có ý nghĩa như sau:

- **SELECTEDVALUE(dimCategory[Category]):** Hàm này sẽ trả về giá trị duy nhất đã được chọn trong cột Category của bảng dimCategory. Nếu có nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **ISBLANK(SELECTEDVALUE(dimCategory[Category])):** Hàm ISBLANK kiểm tra xem giá trị trả về từ SELECTEDVALUE(dimCategory[Category]) có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.

- **IF(ISBLANK(SELECTEDVALUE(dimCategory[Category])), "All Categories", SELECTEDVALUE(dimCategory[Category])):** Hàm IF sẽ kiểm tra điều kiện đầu tiên. Nếu điều kiện đúng (TRUE), nó sẽ trả về “All Categories”. Nếu điều kiện sai (FALSE), nó sẽ trả về giá trị đã được chọn trong cột Category.

4.1.6. dimSubcategory

- a. Measure hiển thị Subcategory được chọn

```
1 SelectedSubcategories = IF(ISBLANK(SELECTEDVALUE(dimSubcategory[Subcategory])), "All SubCategories", SELECTEDVALUE(dimSubcategory[Subcategory]))
```

Hình 88 - DAX hiển thị chủ đề cụ thể được chọn

Biến **SelectedSubcategories** sẽ chứa giá trị “All SubCategories” nếu không có giá trị nào được chọn trong cột Subcategory, hoặc nó sẽ chứa giá trị đã được chọn. Cụ thể các hàm bên trong measure có ý nghĩa như sau:

- **SELECTEDVALUE(dimSubcategory[Subcategory]):** Hàm này sẽ trả về giá trị duy nhất đã được chọn trong cột Subcategory của bảng dimSubcategory. Nếu có nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **ISBLANK(SELECTEDVALUE(dimSubcategory[Subcategory])):** Hàm ISBLANK kiểm tra xem giá trị trả về từ SELECTEDVALUE(dimSubcategory[Subcategory]) có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.
- **IF(ISBLANK(SELECTEDVALUE(dimSubcategory[Subcategory])), "All SubCategories", SELECTEDVALUE(dimSubcategory[Subcategory])):** Hàm IF sẽ kiểm tra điều kiện đầu tiên. Nếu điều kiện đúng (TRUE), nó sẽ trả về “All SubCategories”. Nếu điều kiện sai (FALSE), nó sẽ trả về giá trị đã được chọn trong cột Subcategory.

- b. Measure xếp hạng các Subcategory theo lượt xem trung bình

```
1 AverageViewsRanking =
2 IF (
3   ISINSCOPE ( 'dimSubcategory'[Subcategory] ),
4   RANKX (
5     ALLSELECTED ( dimSubcategory[Subcategory] ),
6     CALCULATE ( AVERAGE(factGYS[Video Views]) ),
7     ,
8     DESC,
9     DENSE
10   )
11 )
```

Hình 89 - DAX xếp hạng chủ đề cụ thể theo lượt xem trung bình

Biến **AverageViewsRanking** sẽ chứa xếp hạng của các Subcategory dựa trên lượt xem trung bình của video, nếu Subcategory đang trong phạm vi của một bộ lọc.

- **ISINSCOPE ('dimSubcategory'[Subcategory])**: Hàm ISINSCOPE kiểm tra xem cột Subcategory của bảng dimSubcategory có đang trong phạm vi của một bộ lọc hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.
- **ALLSELECTED (dimSubcategory[Subcategory])**: Hàm ALLSELECTED trả về tất cả các giá trị đã được chọn trong cột Subcategory của bảng dimSubcategory, bao gồm cả các bộ lọc hiện tại.
- **CALCULATE (AVERAGE(factGYS[Video Views]))**: Hàm CALCULATE thay đổi ngữ cảnh của cột Video Views trong bảng factGYS để tính toán giá trị trung bình.
- **RANKX (ALLSELECTED (dimSubcategory[Subcategory]), CALCULATE (AVERAGE(factGYS[Video Views])), , DESC, DENSE)**: Hàm RANKX sẽ xếp hạng các giá trị đã được chọn trong cột Subcategory dựa trên giá trị trung bình của Video Views. Nó sắp xếp theo thứ tự giảm dần (DESC) và sử dụng phương pháp xếp hạng dày đặc (DENSE), tức là không có “khoảng trống” giữa các xếp hạng.
- **IF (ISINSCOPE ('dimSubcategory'[Subcategory]), RANKX (ALLSELECTED (dimSubcategory[Subcategory]), CALCULATE (AVERAGE(factGYS[Video Views])), , DESC, DENSE))**: Hàm IF sẽ kiểm tra điều kiện đầu tiên. Nếu điều kiện đúng (TRUE), nó sẽ trả về xếp hạng được tính toán bởi RANKX.

4.1.7. dimTypeofYoutuber

```
TypeofYoutubers = IF(ISBLANK(SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber])), "All Type of Youtuber", SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber]))
```

Hình 90 - DAX hiển thị loại kênh Youtube được chọn

Biến TypeofYoutubers sẽ trả về “All Type of Youtuber” nếu không có giá trị nào được chọn trong cột TypeofYoutuber, hoặc nó sẽ chứa giá trị đã được chọn là: Individual/Group/ Company. Phân tích các hàm sử dụng trong measure:

- **SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber])**: Hàm này sẽ trả về giá trị duy nhất đã được chọn trong cột TypeofYoutuber của bảng dimTypeofYoutuber. Nếu có nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **ISBLANK(SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber]))**: Hàm ISBLANK kiểm tra xem giá trị trả về từ SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber]) có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.

- **IF(ISBLANK(SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber])), "All Type of Youtuber", SELECTEDVALUE(dimTypeofYoutuber[TypeofYoutuber])):** Hàm IF sẽ kiểm tra điều kiện đầu tiên. Nếu điều kiện đúng (TRUE), nó sẽ trả về “All Type of Youtuber”. Nếu điều kiện sai (FALSE), nó sẽ trả về giá trị đã được chọn trong cột TypeofYoutuber.

4.1.8. dimYoutuberCreatorAwards

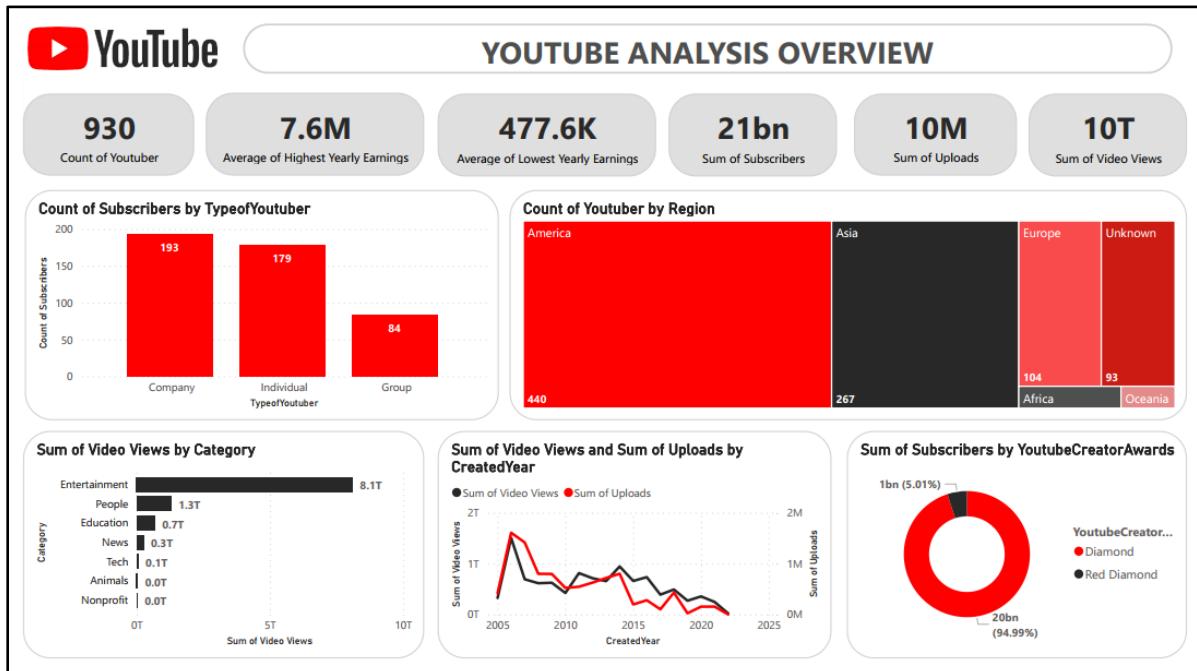
```
Awards = IF(ISBLANK(SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards])), "All Youtuber Creator Awards", SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards]))
```

Hình 91 - DAX hiển thị loại giải thưởng kênh được chọn

Measure hiển thị Youtuber Creator Awards được chọn. Biến Awards sẽ trả về “All Youtuber Creator Awards” nếu không có giá trị nào được chọn trong cột YoutubeCreatorAwards, hoặc nó sẽ chứa giá trị đã được chọn là: Diamond / Red Diamond. Phân tích các hàm sử dụng trong measure:

- **SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards]):** Hàm này sẽ trả về giá trị duy nhất đã được chọn trong cột YoutubeCreatorAwards của bảng dimYoutubeCreatorAwards. Nếu có nhiều hơn một giá trị được chọn, hàm này sẽ trả về BLANK().
- **ISBLANK(SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards])):** Hàm ISBLANK sẽ kiểm tra xem giá trị trả về từ SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards]) có phải là BLANK() hay không. Nếu có, nó trả về TRUE, nếu không, nó trả về FALSE.
- **IF(ISBLANK(SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards])), "All Youtube Creator Awards", SELECTEDVALUE(dimYoutubeCreatorAwards[YoutubeCreatorAwards])):** Hàm IF sẽ kiểm tra điều kiện đầu tiên. Nếu điều kiện đúng (TRUE), nó sẽ trả về “All Youtuber Creator Awards”. Nếu điều kiện sai (FALSE), nó sẽ trả về giá trị đã được chọn trong cột YoutubeCreatorAwards.

4.2. Phân tích tổng quan về hiệu suất của các kênh Youtube (Overview)



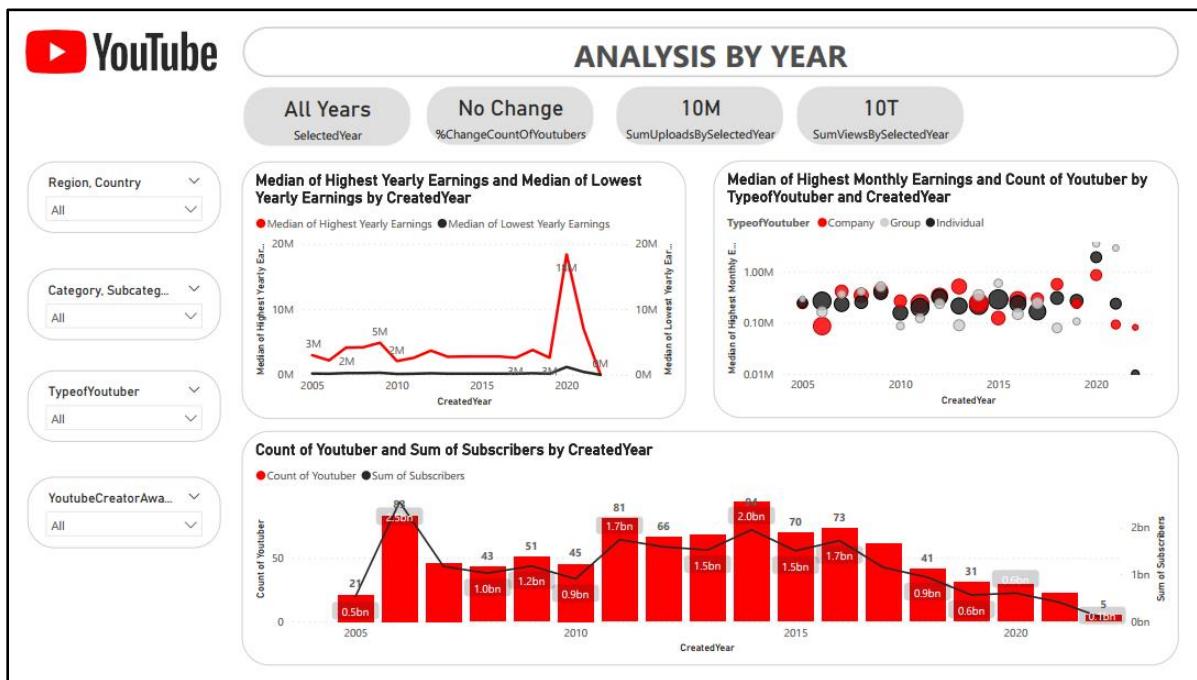
Hình 92 - Trang phân tích tổng quan hiệu suất kênh Youtube

Từ bộ dữ liệu thu thập các thông tin về Youtube, nền tảng chia sẻ video hàng đầu thế giới, có tổng cộng 930 YouTuber đang tạo ra nội dung. Họ kiếm được tiền từ video do chính mình tạo ra, với thu nhập trung bình hàng năm cao nhất có thể đạt đến mức 7.6 triệu đô la, thậm chí mức trung bình thu nhập hàng năm thấp nhất cũng lên đến 477 nghìn đô la. Điều này không chỉ cho thấy sức hút mạnh mẽ của nền tảng này, mà còn phản ánh tiềm năng kiếm tiền từ việc sáng tạo nội dung. Tổng số người đăng ký trên tất cả các kênh lên đến 21 tỷ, một con số đáng kinh ngạc phản ánh sức ảnh hưởng to lớn của nền tảng này đối với cuộc sống hàng ngày của chúng ta. Với 10 triệu video được tải lên và tổng cộng 10 nghìn tỷ lượt xem, Youtube không chỉ là một nền tảng chia sẻ video mà còn là một cơ hội để mọi người thể hiện tài năng và kiếm tiền từ đam mê của mình.

Các công ty đang chiếm ưu thế trên Youtube, nhờ nguồn lực lớn và thương hiệu của họ, cho phép tạo và quảng bá nội dung chất lượng. Tuy nhiên, các YouTuber cá nhân và nhóm YouTuber vẫn đóng góp một phần quan trọng vào thế giới YouTube, tạo ra nội dung đa dạng và phong phú hàng ngày. Với số lượng YouTuber đông đảo, Châu Mỹ trở thành cái nôi của những người sáng tạo nội dung, định hình xu hướng và tạo ra những hiện tượng văn hóa mới. Mảng giải trí là lĩnh vực mà người xem quan tâm nhiều nhất, phản ánh nhu cầu giải trí không ngừng của con người. YouTube trở thành sân khấu toàn cầu cho mọi người thể hiện và chia sẻ niềm đam mê.

Từ năm 2010 trở đi, người dùng YouTube đã chuyển hướng sự quan tâm của mình từ việc tải video lên sang việc xem video. Điều này có thể được giải thích bởi sự gia tăng của nội dung chất lượng cao từ các nhà sản xuất chuyên nghiệp và sự thay đổi trong hành vi của người dùng khi họ dành nhiều thời gian hơn để xem và khám phá nội dung thay vì tạo ra nội dung của riêng mình.Thêm vào đó, những người sáng tạo nội dung xuất sắc nhận được giải thưởng Diamond và Red Diamond, với tổng số người đăng ký của các kênh này lên đến hơn 20 tỷ, cho thấy các Youtuber trong bộ dữ liệu này hầu như đều là các Youtuber có sự ảnh hưởng lớn và thành công trên nền tảng này.

4.3. Phân tích hiệu suất của các kênh Youtube dựa theo năm (Analysis by Year)



Hình 93 - Phân tích hiệu suất của các kênh Youtube theo Year

Những “ngôi sao” Youtube được sinh ra khi nào? Xu hướng của người dùng

Nhìn vào biểu đồ cột, có thể thấy sự tăng trưởng về số lượng các kênh và nhà sáng tạo nội dung hàng đầu của nền tảng YouTube không tuân theo 1 xu hướng tuyến tính nào mà biến động qua từng thời kỳ. Có thể xem 2006 - 1 năm kể từ lúc YouTube được ra mắt, là một cột mốc đặc biệt đánh dấu sự khởi đầu của gần 9% các “gã khổng lồ” trên YouTube. Bên cạnh đó, giai đoạn từ năm 2011-2016 là thời kỳ chứng kiến sự ra đời của phần lớn những “ngôi sao” YouTube được thống kê trong bộ dữ liệu. Trong đó 2014 là đỉnh điểm khi có tới hơn 10% các kênh được thành lập vào năm này. Kể từ sau 2017, tỷ lệ số kênh và nhà sáng tạo dẫn đầu được thành lập vào

thời gian này giảm dần đều.

Sự phổ biến của Youtube là một điều không thể phủ nhận, nhìn vào sự tăng trưởng tổng số người đăng ký (subscribers) của những kênh thành lập vào năm 2006 có thể cảm nhận được sự hưởng ứng và hứng thú của cộng đồng người dùng Internet dành cho nền tảng. Sau 5 năm nền tảng phát triển và hoạt động, có lẽ các nhà sản xuất nội dung đã thật sự đạt được sự trưởng thành trong định hướng và chất lượng của nội dung mà họ đăng tải, từ đây bắt đầu thời kỳ rực rỡ khi mà rất nhiều kênh Youtube đạt hàng trăm triệu lượt đăng ký ra đời vào thời gian này.

2014 là năm đứng đầu về số lượng kênh được thành lập (trong phạm vi của bộ dữ liệu), tương ứng với tổng số lượt đăng ký của các kênh này là hơn 1.9 tỷ lượt, chỉ xếp sau 2006 với tổng hơn 2.5 tỷ người đăng ký kênh (tính theo thời điểm thống kê). Từ đây có thể suy luận rằng những kênh Youtube lâu đời thu hút được đông đảo khán giả trung thành. Nguyên nhân có thể đến từ lợi thế quá trình xây dựng kênh lâu dài, chất lượng nội dung của các kênh, hay đến từ chính hành vi, xu hướng của người dùng Youtube.

Sau 2017 có thể gọi là thời điểm bão hòa khi mà số lượng kênh thuộc top đầu và tổng subscribe của các kênh thành lập trong thời điểm này tuột dốc. Một phần có thể là vì tuổi đời non trẻ của các kênh so với những “gã khổng lồ” sinh ra từ các thập kỷ trước. Bên cạnh đó, việc ngày càng nhiều người tham gia vào sân chơi sáng tạo nội dung khiến cho sự cạnh tranh trở nên khốc liệt hơn. Trải qua một quá trình hoạt động lâu dài, người dùng Youtube cũng hình thành nhiều tiêu chuẩn và chọn lọc đối với các nội dung họ tiếp cận. Một yếu tố nữa có thể là nguyên nhân của hiện tượng này, chính là sự ra đời của rất nhiều nền tảng mạng xã hội, nền tảng chia sẻ nội dung khác. Nó khiến cho cả khán giả lẫn người sản xuất nội dung cắt giảm thời gian và tài nguyên của mình dành cho Youtube.

Nhìn chung, Youtube đã có một sự phát triển nhanh chóng về người dùng của nền tảng ngay từ những ngày đầu; sau thời kỳ hoàng kim, Youtube đang trải qua sự thay đổi trong xu hướng của ngành sáng tạo nội dung và lĩnh vực phát triển phương tiện truyền thông xã hội.

Thu nhập của những kênh Youtube

Với biểu đồ đường phản ánh thu nhập điển hình hàng năm cao nhất (Highest

Yearly Earnings), có thể thấy nhiều biến động qua các năm thành lập kênh, với một số năm có sự tăng trưởng mạnh mẽ và những năm khác có sự giảm sút.

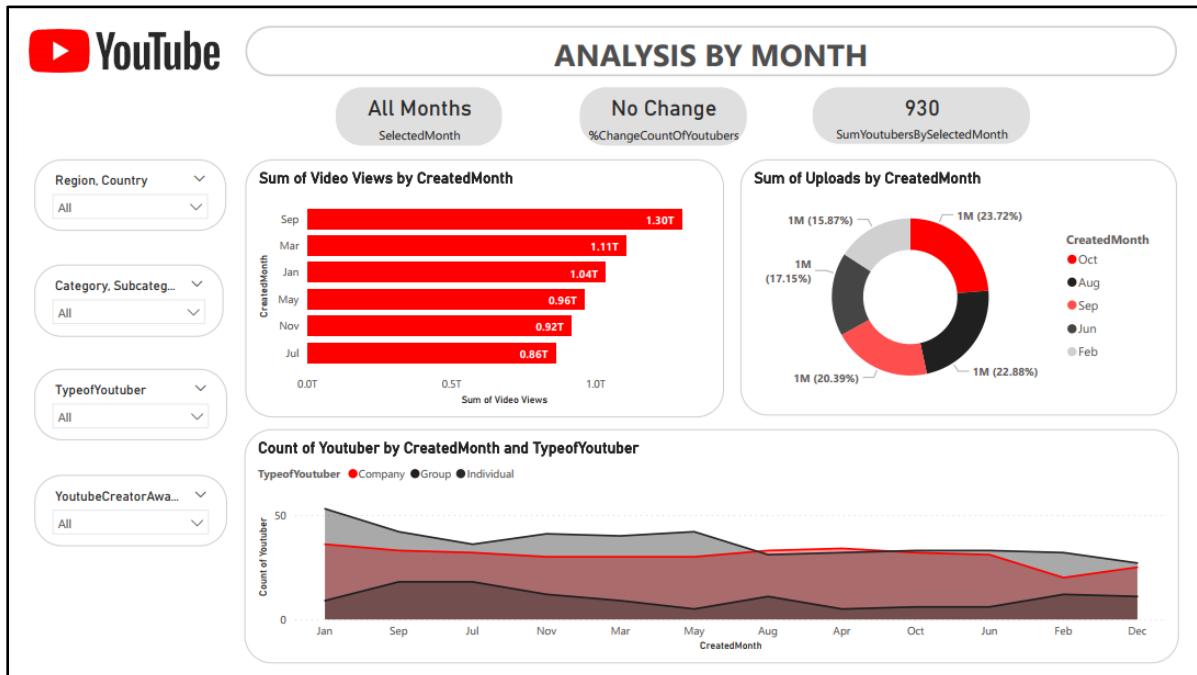
Thu nhập hằng năm đa phần dao động từ khoảng 2-5 triệu đô. Trung bình thu nhập cao nhất của những kênh thành lập từ 2022 là bé nhất khi chưa đến 500 đô. Nguyên nhân có thể là do đây là những kênh mới, còn non trẻ do đó khả năng tối đa lợi nhuận kiếm được còn chưa tốt. Trong khi đó, những kênh thành lập vào năm 2020 đạt thu nhập tăng vọt khi trung bình thu nhập cao nhất của những kênh Youtube đạt đỉnh điểm là hơn 18 triệu đô. Thu nhập đỉnh điểm này có thể phản ánh sự thay đổi trong hành vi người dùng hoặc các yếu tố thị trường khác, như sự gia tăng nhu cầu giải trí trực tuyến tăng lên đáng kể do đại dịch COVID-19. Thuật toán của nền tảng cũng cần được cân nhắc bởi những nền tảng này thường cập nhật thuật toán, điều này ảnh hưởng đến lượng người xem, khả năng tiếp cận khán giả và thu nhập của các kênh.

Nhiều người cho rằng tiềm năng kiếm tiền từ YouTube là rất lớn, để đánh giá nhận định này, ta sẽ tiếp tục xem xét trung bình thu nhập hằng năm thấp nhất của những kênh Youtube. Có thể thấy chênh lệch giữa mức thu nhập cao nhất và thấp nhất tương đối lớn. Vậy không thể cho rằng thu nhập từ YouTube không phải lúc nào cũng ổn định và nó có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau.

Ngoài ra, quy mô nhân sự đơn vị đứng sau kênh sản xuất nội dung cũng có thể là một yếu tố ảnh hưởng thu nhập của kênh Youtube. Với biểu đồ phân tán, dễ dàng nhận thấy Company và Individual chiếm ưu thế hơn về tỷ lệ. Tuy nhiên, trung bình thu nhập hàng tháng của các phân loại Youtuber biến động phức tạp, không tuân theo bất kỳ quy tắc nào. Nếu những công ty sản xuất nội dung thành lập từ những năm 2018 về trước có xu hướng dẫn đầu về thu nhập, thì kênh thành lập những năm gần đây đã có sự thay đổi về xếp hạng. Kênh các nhân hoặc kênh với quy mô nhỏ như nhóm tác giả lần lượt vượt mặt các công ty về trung bình thu nhập hàng tháng. Điều này phản ánh sự cạnh tranh gia tăng và khó khăn trong việc kiếm tiền từ YouTube cũng như xu hướng vận động khó lường của nền tảng này.

Nhìn chung, kiếm tiền từ Youtube luôn tồn tại cơ hội và thách thức. Cần tìm hiểu rõ nhiều yếu tố cùng với sự ảnh hưởng của các sự kiện toàn cầu và xu hướng thị trường.

4.4. Phân tích hiệu suất của các kênh Youtube dựa theo tháng (Analysis by Month)



Hình 94 - Phân tích hiệu suất của các kênh Youtube theo Month

Từ biểu đồ đường, nhận thấy các kênh Youtube trong bộ dữ liệu được thành lập nhiều nhất vào tháng 1 và thấp nhất vào tháng 12. Trong đó, các kênh Youtube thuộc phân loại Individual là chiếm tỷ lệ cao nhất, xếp thứ 2 là Company và cuối cùng là Group. Có 2 tháng mà số lượng kênh Company được lập ra nhỉnh hơn kênh Individual là tháng 8 và tháng 4.

Thông qua biểu đồ top 6 tháng thành lập kênh có tổng lượt xem cao nhất (views), có thể thấy chỉ số này của các kênh phân bố không đều giữa các tháng thành lập kênh. Trong đó tổng lượt xem của những kênh thành lập vào tháng 9 là cao nhất đạt gần 1.3 nghìn tỷ views. Điều này là dễ hiểu vì tháng thành lập kênh ít có ý nghĩa ảnh hưởng đến hiệu suất kênh.

Với biểu đồ tròn, top 5 tháng lập kênh có tổng video đăng tải lớn nhất lần lượt là tháng 10, tháng 8, tháng 9, tháng 6 và tháng 2. Tỷ lệ của các tháng này tương đối đồng đều, trong đó tháng 10 chiếm phần trăm nhất với hơn 1.3 triệu video (>23%), và tháng 2 chiếm thấp nhất với hơn 900 nghìn video (>15%).

Nhìn vào biểu đồ cột, có thể thấy sự tăng trưởng về số lượng các kênh và nhà sáng tạo nội dung hàng đầu của nền tảng Youtube không tuân theo 1 xu hướng tuyến tính nào mà biến động qua từng thời kỳ.

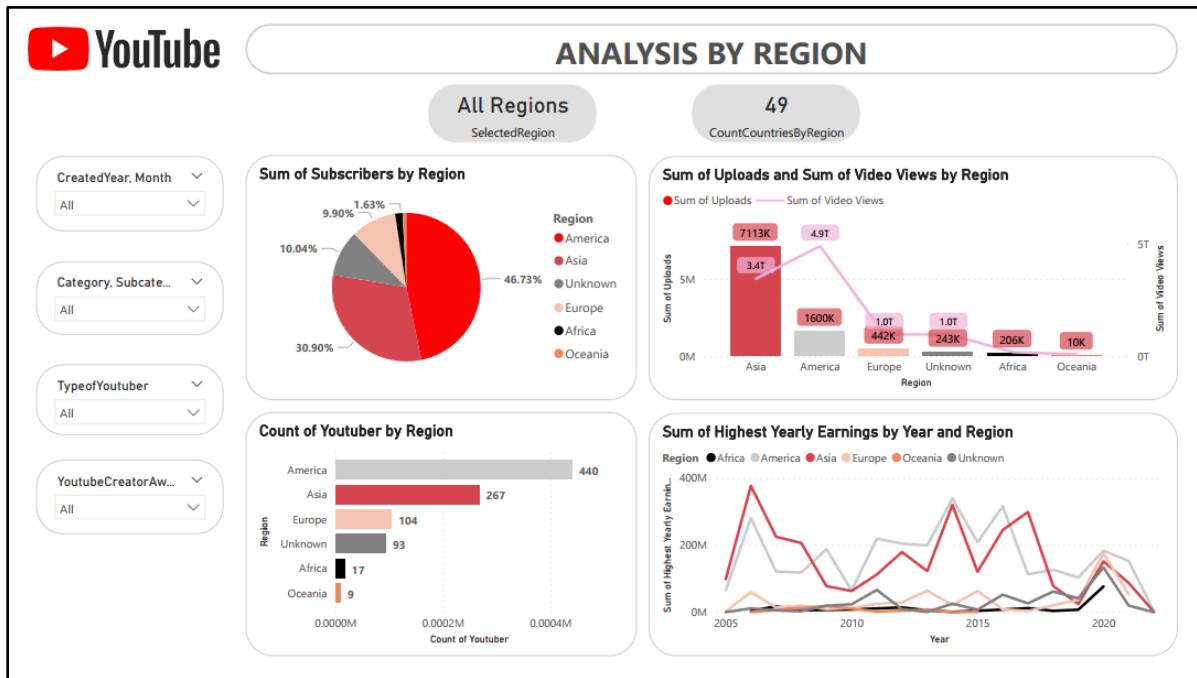
- Năm 2006, cột mốc đánh dấu 1 năm kể từ lúc Youtube được ra mắt, có 9% các kênh hàng đầu được thành lập. Giai đoạn từ năm 2011-2016 là thời kỳ chứng kiến sự ra đời của phần lớn những “ngôi sao” Youtube được thống kê trong bộ dữ liệu, với đỉnh điểm là năm 2014 khi hơn 10% kênh được lập. Kể từ sau 2017, tỷ lệ số kênh và nhà sáng tạo dần đầu được thành lập vào thời gian này giảm dần đều.
- Nhìn vào 2006 có thể cảm nhận được sự hưởng ứng và hứng thú của cộng đồng người dùng Internet dành cho nền tảng. Sau 5 năm, có lẽ các nhà sản xuất nội dung đã thật sự đạt được sự trưởng thành trong định hướng và chất lượng của nội dung mà họ đăng tải, từ đây bắt đầu thời kỳ rực rỡ khi mà rất nhiều kênh Youtube đạt hàng trăm triệu lượt đăng ký ra đời vào thời gian này. Thời điểm bão hòa sau 2017, số lượng kênh thuộc top đầu và tổng subscribe của các kênh thành lập trong thời điểm này tuột dốc. Một phần có thể là vì tuổi đời non trẻ của các kênh mới, một phần do ngày càng nhiều người tham gia vào sân chơi sáng tạo nội dung khiến cho sự cạnh tranh trở nên khốc liệt hơn. Trải qua một quá trình hoạt động lâu dài, người dùng Youtube cũng hình thành nhiều tiêu chuẩn và chọn lọc đối với các nội dung họ tiếp cận. Một yếu tố nữa có thể là nguyên nhân của hiện tượng này, chính là sự ra đời của rất nhiều nền tảng mạng xã hội, nền tảng chia sẻ nội dung khác.
- Thu Nhập Của Kênh YouTube: Thu nhập hàng năm dao động từ 2-5 triệu đô. Trung bình thu nhập cao nhất của những kênh thành lập từ 2022 là bé nhất khi chưa đến 500 đô. Nguyên nhân có thể là do đây là những kênh mới, khả năng tối đa lợi nhuận kiếm được còn chưa tốt. Trong khi đó, những kênh thành lập vào năm 2020 đạt thu nhập tăng vọt khi trung bình thu nhập cao nhất của những kênh Youtube đạt đỉnh điểm là hơn 18 triệu đô, phản ánh ảnh hưởng của COVID-19 và thay đổi thuật toán.

Xu Hướng Thành Lập Kênh qua các Tháng: Các kênh được lập nhiều nhất vào tháng 1 và ít nhất vào tháng 12, với sự thay đổi trong hiệu suất và thu nhập theo thời gian.

- Tổng lượt xem của những kênh thành lập vào tháng 9 là cao nhất đạt gần 1.3 nghìn tỷ views.
- Với biểu đồ tròn, top 5 tháng lập kênh có tổng video đăng tải lớn nhất lần lượt là tháng 10, tháng 8, tháng 9, tháng 6 và tháng 2. Tỷ lệ của các tháng này tương đối đồng đều, trong đó tháng 10 chiếm phần trăm nhất với hơn 1.3 triệu video (>23%), và tháng 2 chiếm thấp nhất với hơn 900 nghìn video (>15%).

- Những biến động này tương đối ít ý nghĩa vì tháng thành lập kênh có ít ảnh hưởng đến hiệu suất của kênh.

4.5. Phân tích hiệu suất của các kênh Youtube dựa theo vùng (Analysis by Region)



Hình 95 - Phân tích hiệu suất của các kênh Youtube theo Region

Đầu tiên nói về pie chart cho thấy tổng số lượt đăng ký ở mỗi khu vực. Điều đáng chú ý là Châu Mỹ dẫn đầu với hơn 10 tỷ đăng ký, chiếm 46,73%, nhiều hơn 16% so với Châu Á. Châu Mỹ thu hút nhiều lượt đăng ký hơn cả có thể do nội dung của họ thu hút hơn, cùng với đó, YouTube là nền tảng được tạo ra tại Mỹ nên nó được mọi người ở đây tin dùng nhiều hơn. Bên cạnh đó Châu Mỹ là nơi đa dạng chủng tộc, đa phần mọi người ở đây giao tiếp bằng những loại ngôn ngữ phổ biến trên thế giới. Điều này làm nội dung của họ thu hút nhiều người xem hơn. . Những điều trên giúp Châu Mỹ tạo ra thị trường tiềm năng cho lớn cho các nhà tiếp thị và nhà sản xuất nội dung.接下來，我們將分析不同地區的YouTuber數量。美國有440位YouTuber，亞洲有267位，歐洲有104位，未知地區有93位，非洲有17位，大洋洲有9位。

Tiếp theo, chúng ta nhìn vào biểu đồ cột đầu tiên, thể hiện tổng số YouTuber theo vùng. Một lần nữa, Châu Mỹ vẫn đứng đầu với số lượng YouTuber lớn nhất.. Chúng tôi rằng YouTube đã trở thành một nền tảng quan trọng cho việc chia sẻ và tiếp

cận nội dung tại Châu Mỹ. Đồng thời, biểu đồ cột thứ hai cho thấy tổng số lượt xem video theo vùng. Một lần nữa, Châu Mỹ dẫn đầu với số lượt xem video cao nhất. Trong khi đó số lượng video được upload của Châu Á lớn hơn gấp 6 lần Châu Mỹ, nhưng lượt xem của Châu Mỹ lại gấp 1.36 lần Châu Á, và bỏ xa các khu vực còn lại. Điều này cho thấy rằng nội dung của các YouTuber ở Châu Mỹ không chỉ thu hút được sự quan tâm của cộng đồng người dùng trong khu vực, mà còn thu hút được sự quan tâm từ người dùng ở các khu vực khác. Cũng qua đó ta có thể thấy youtuber ở Châu Á chăm chỉ sản xuất video hơn các khu vực còn lại khá nhiều.

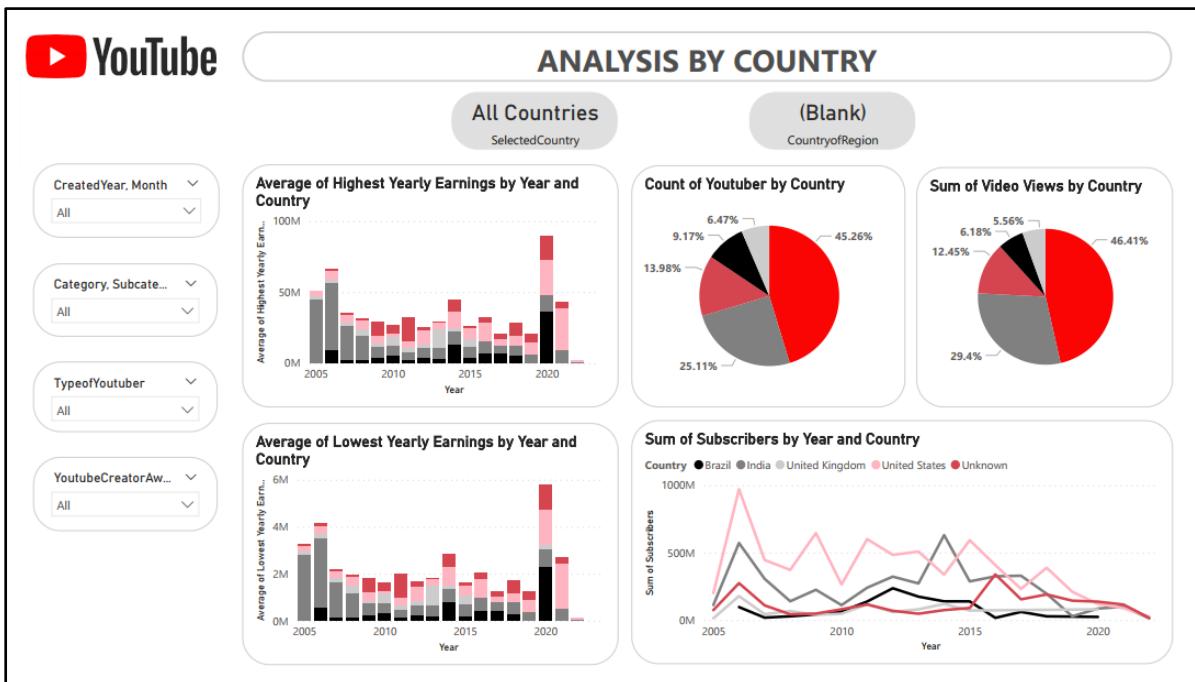
Đi sâu vào số liệu của những khu vực còn lại, chúng ta có thể thấy rõ họ có ít sự quan tâm với youtube hơn hẳn châu Mỹ và châu Á. Trước tiên nói về châu Phi, ở đây tình hình kinh tế và đời sống còn bấp bênh, internet và các thiết bị điện tử không phổ biến như 4 khu vực còn lại. Thêm nữa châu Phi ít có những Youtuber có lượt theo dõi hàng đầu trong bộ dữ liệu này cũng là điều dễ hiểu. Bàn về khu vực châu Âu, ở đây các số liệu của họ đứng thứ ba toàn cầu nhưng chi tiết thì số liệu luôn thấp hơn châu Mỹ và châu Á ít nhất là 2 lần. Nếu dùng hai lý do của châu Mỹ và châu Á ta sẽ có thể thấy rằng, châu Âu là khu vực có diện tích nhỏ nhưng lại có nhiều quốc gia, và mỗi quốc gia lại có ngôn ngữ riêng. Có những quốc gia có ngôn ngữ không phổ biến và khó học nên nếu dùng ngôn ngữ mẹ đẻ làm video thì họ sẽ bị hạn chế với quốc tế. So về dân cư, khu vực này cũng có số dân cư ít ỏi và dân số già có chiều hướng gia tăng nên youtube có thể ít thu hút họ hơn. Cũng với lý do dân số ta có thể thấy được châu Úc luôn đứng cuối các bảng xếp hạng trên biểu đồ. Vì dân số còn ít hơn cả châu Âu nên số lượng Youtuber của họ cũng ít và số lượng Youtuber có số lượng người đăng ký hàng đầu lại càng ít. Điều này dẫn đến các số liệu khác của họ luôn thua xa tất cả khu vực khác.

Cuối cùng bàn về tổng thu nhập cao nhất từng năm của mỗi vùng. Ở giai đoạn đầu Châu Á thu nhập cao hơn, dao động từ 200 triệu đô la Mỹ cho đến hơn 370 triệu đô, tuy nhiên từ giai đoạn 2009 trở về sau, Châu Mỹ vẫn tiếp tục dẫn đầu với thu nhập cao nhất. Điều này cũng tương ứng với lượt xem lớn hơn của Châu Mỹ so với Châu Á, nhưng số liệu cũng không quá chênh lệch. Những châu lục còn lại với lượt đăng ký, số lượng youtuber, số video được upload đều ít hơn nhiều so với Châu Mỹ và Châu Á, cho nên thu nhập của họ ít hơn hẳn hai khu vực này, cao nhất cũng chỉ có hơn 60 triệu đô.

Nhìn chung hai thị trường Châu Mỹ và Châu Á hoạt động sôi nổi trên youtube

hơn rất nhiều so với những khu vực khác. Kéo theo đó là sự cạnh tranh gay gắt giữa những Youtuber này.

4.6. Phân tích hiệu suất của các kênh Youtube dựa theo quốc gia (Analysis by Country)



Hình 96 - Phân tích hiệu suất của các kênh Youtube theo Country

YouTube trở thành một nền tảng phổ biến toàn cầu với những người sáng tạo nội dung ở tất cả các quốc gia.

Có thể thấy Hoa Kỳ dẫn đầu với gần phân nửa số lượng Youtuber (chiếm tận 45%). Ấn Độ cũng không kém cạnh khi chiếm $\frac{1}{4}$ pie chart. Cùng với đó, lượt views của video cũng tỷ lệ thuận với số lượng youtuber hiện có. Những nước có lượt view cao hàng đầu lần lượt là Hoa Kỳ với 46% (hơn 4 triệu tỷ views), Ấn Độ với gần 30% (hơn 2 triệu tỷ views), Brazil và Vương quốc Anh.

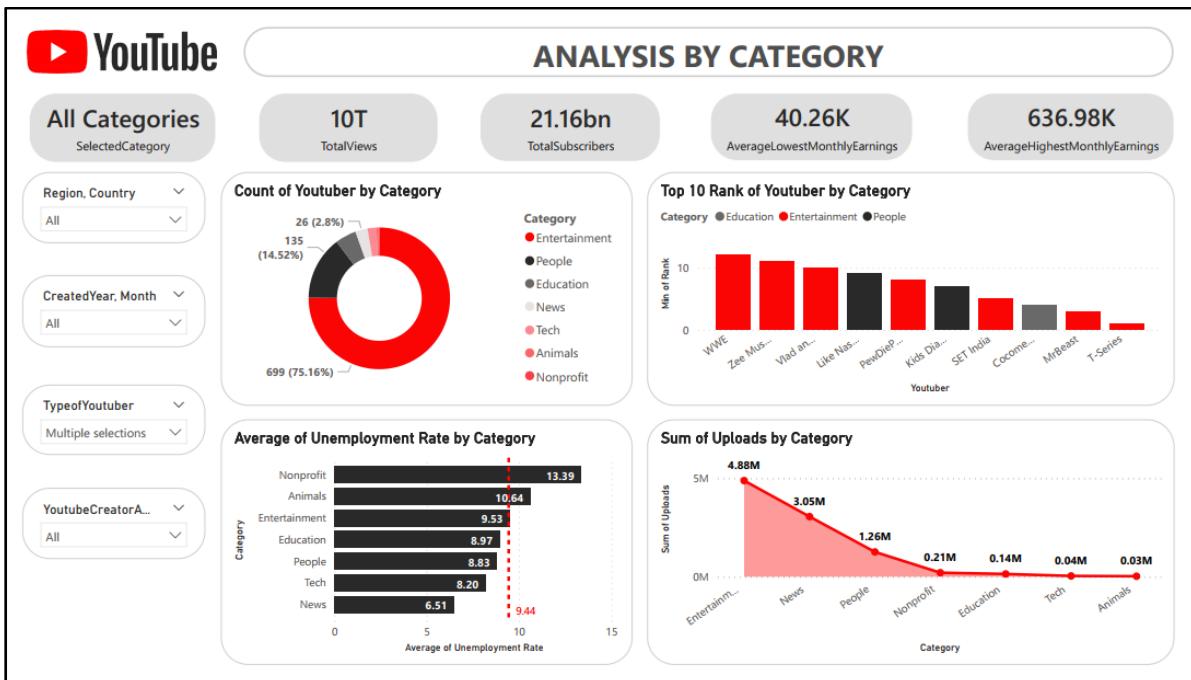
Về sự phát triển của các kênh youtube từ 2005 đến nay thì biểu đồ dạng line “Tổng số lượng người đăng ký của các quốc gia” cho thấy sự dao động tăng giảm theo từng năm mà không phải chỉ là xu hướng đi lên hoặc chỉ đi xuống. Các quốc gia đều có những con đường riêng biệt, phản ánh sự đa dạng trong hành vi tạo và tiêu thụ nội dung. Mỹ vẫn luôn đứng đầu và có sự dao động mạnh hơn các nước còn lại. Điểm đặc biệt là ở năm 2006 lượt người đăng ký tăng vượt bậc. Theo dòng lịch sử, năm 2006 Google chính thức mua lại YouTube với giá 1,65 tỷ đô la Mỹ làm cho YouTube phát triển vượt bậc. Cùng năm đó YouTube đoạt giải "Phát minh năm 2006" cũng tạo tiếng

vang không nhỏ, góp phần thu hút rất nhiều lượt người tham gia và đăng ký các kênh youtube. Điều này giúp giúp số liệu của Mỹ tăng vọt vì quốc gia này là nơi tạo ra Youtube và đặt trụ sở chính ở đây. Vào giai đoạn 2010-2015, internet phát triển, số lượt người đăng ký youtube theo đó cũng tăng, nhất là ở Ấn Độ có bước tăng vọt vào năm 2014. Trái ngược với đó thì Mỹ lại không có quá nhiều sự biến động, có thể là do chính đất nước của họ đã tạo ra nhiều ứng dụng mạng xã hội hơn và nó đã cạnh tranh với Youtube. Ba thị trường còn lại ít biến động hơn hẳn và mức số liệu khá khiêm tốn so với Mỹ và Ấn Độ. Những năm sau đó số lượt người đăng ký giảm dần, có thể là do thị trường xuất hiện thêm các nền tảng khác cạnh tranh mạnh, có thể kể đến đối thủ lớn nhất của Youtube hiện nay là Tik Tok.

Cuối cùng bàn về thu nhập bình quân của các kênh youtube theo từng năm của mỗi quốc gia. Ở đây ta thấy sự ổn định số liệu trong giai đoạn 2008-2019. Điều đặc biệt ở đây là thu nhập của youtuber Ấn Độ hai năm 2005 và 2006 áp đảo tất cả, thậm chí gấp nhiều lần Mỹ - quốc gia có số youtuber và lượt người đăng ký gấp đôi họ. Nhưng điều bất thường ở đây là khi giai đoạn Internet bùng nổ như đã phân tích ở trên, mặc dù số người đăng ký tăng vọt nhưng thu nhập của họ lại không bằng khoảng thời gian ban đầu khi Youtube thành lập. Có lẽ lúc vì sự phổ biến của Internet nên số người mới gia nhập youtube nhiều hơn, dẫn theo đó là sự cạnh tranh khốc liệt, số tiền kiếm được của các Youtuber ở top đầu số người đăng ký được chia cho những Youtuber mới nổi. Ngoài ra năm 2020 thu nhập cao đột biến có thể là do đại dịch Covid-19 khiến nhiều người ở nhà và xem youtube nhiều. Đặc biệt là ở Brasil và Mỹ năm 2020 họ chịu thiệt hại khá nặng nề vì dịch bệnh và phải ở nhà cách ly nhiều tháng liền.

Nhìn chung thì ở dashboard này, các quốc gia có Youtuber ở top đầu đều là quốc gia nói ngôn ngữ phổ biến như tiếng Anh, Bồ Đào Nha, Tây Ban Nha hay những nước đông dân.

4.7. Phân tích hiệu suất của các kênh Youtube dựa theo chủ đề kênh (Analysis by Category)



Hình 97 - Phân tích hiệu suất của các kênh Youtube theo Category

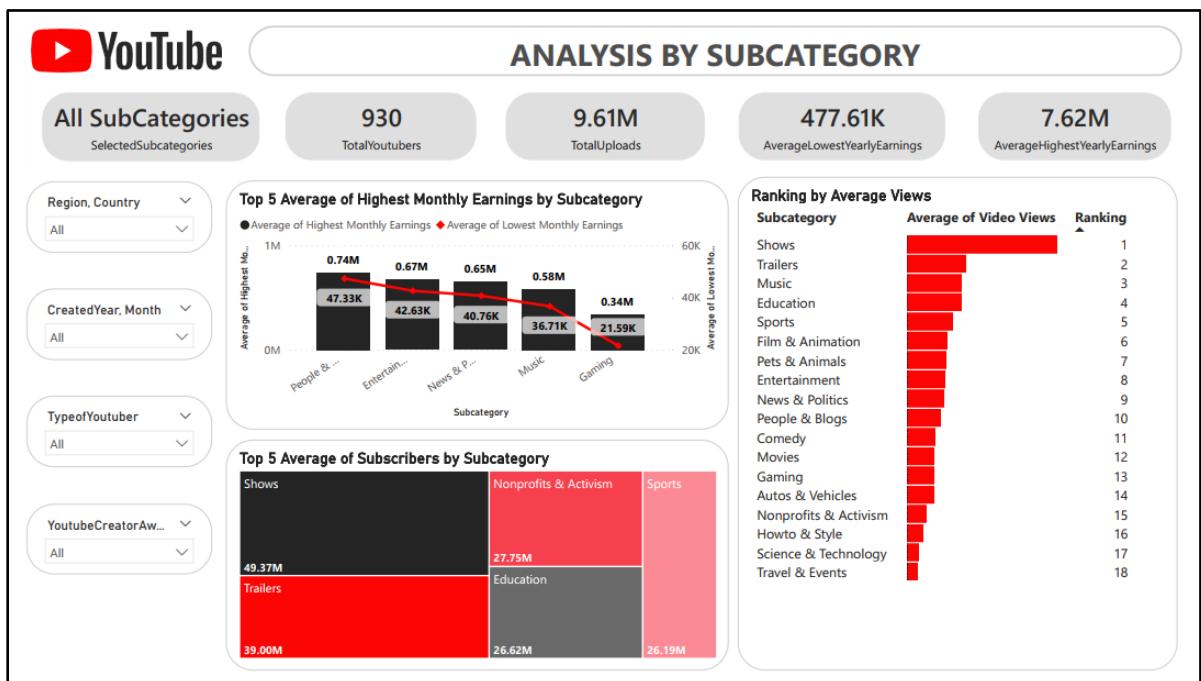
Entertainment là chủ đề phổ biến nhất trong các kênh youtube với hơn 75% youtubers lựa chọn. Theo sau đó là People và Education nhưng số lượng vẫn chênh lệch đáng kể so với Entertainment. Đặc biệt, trong 10 youtuber được xếp hạng cao nhất đã có đến 7 người lựa chọn chủ đề kênh của mình là Entertainment. Bên cạnh đó, chủ đề này cũng chiếm số lượng video đăng tải nhiều nhất với khoảng 4.88 triệu video trong nền tảng Youtube. Tuy nhiên, đây là 1 trong 3 chủ đề có trung bình tỷ lệ thất nghiệp cao nhất. Điều này chứng tỏ các youtuber thường đến từ những quốc gia có tỷ lệ thất nghiệp khá cao. Việc họ lựa chọn chủ đề kênh là Entertainment có thể hiểu được bởi vì đây là chủ đề có phạm vi rộng, liên quan đến nhiều khía cạnh khác nhau như: chương trình thực tế, hài kịch, phim ảnh, âm nhạc,... Bên cạnh đó, chủ đề này đa phần yêu cầu về sự sáng tạo và ít nặng về kiến thức chuyên môn. Do đó, đây là chủ đề được nhiều youtubers lựa chọn thực hiện nhiều nhất.

Hai chủ đề thường được các youtuber lựa chọn xếp sau Entertainment là People và Education với số lượng kênh lần lượt 135 và 45. Trong top 10 youtuber có xếp hạng cao nhất đã có 2 người lựa chọn kênh chủ đề là People và 1 người lựa chọn về Education. Đặc biệt, cả 2 chủ đề này đều có tỷ lệ thất nghiệp trung bình thấp hơn mức trung bình.

- Đối với People, chủ đề này có số lượng video đăng tải đứng thứ 3 với hơn 1 triệu video. Tương tự như Entertainment, People là chủ đề gần gũi, không đòi hỏi quá nhiều kiến thức chuyên môn và liên quan đến nhiều khía cạnh về con người như chia sẻ về cuộc sống, kỹ năng, trải nghiệm,... Vì thế, đây là chủ đề phổ biến thứ 2 sau Entertainment.
- Đối với Education, số lượng video đăng tải khá ít chỉ với khoảng 140 triệu video. Điều này có thể do chủ đề này đòi hỏi nhiều kiến thức và cần có sự chính xác khi chia sẻ. Nhưng cũng không thể phủ nhận sự quan tâm của các youtubers đối với chủ đề này là khá cao.

Kết luận: Entertainment, People và Education là top 3 chủ đề đang được các youtuber hướng đến khi xây dựng kênh youtube của mình. Trong đó Entertainment đang chiếm số lượng áp đảo với hơn 75% youtubers lựa chọn. Tuy nhiên, điều này dẫn đến sự cạnh tranh lớn giữa các youtubers, và gây ra ảnh hưởng về hiệu suất của các kênh youtube như về lượt views, share, subscribes,... Bên cạnh đó, sự cạnh tranh này cũng đặt ra thách thức đối với các youtubers mới tham gia.

4.8. Phân tích hiệu suất của các kênh Youtube dựa theo chủ đề chi tiết của kênh (Analysis by Subcategory)



Hình 98 - Phân tích hiệu suất của các kênh Youtube theo Subcategory

Mặc dù mức độ phổ biến chỉ xếp thứ 2 nhưng People & Blogs lại dẫn đầu về mức trung bình thu nhập hàng tháng cao nhất và thấp nhất. Trung bình thu nhập mà

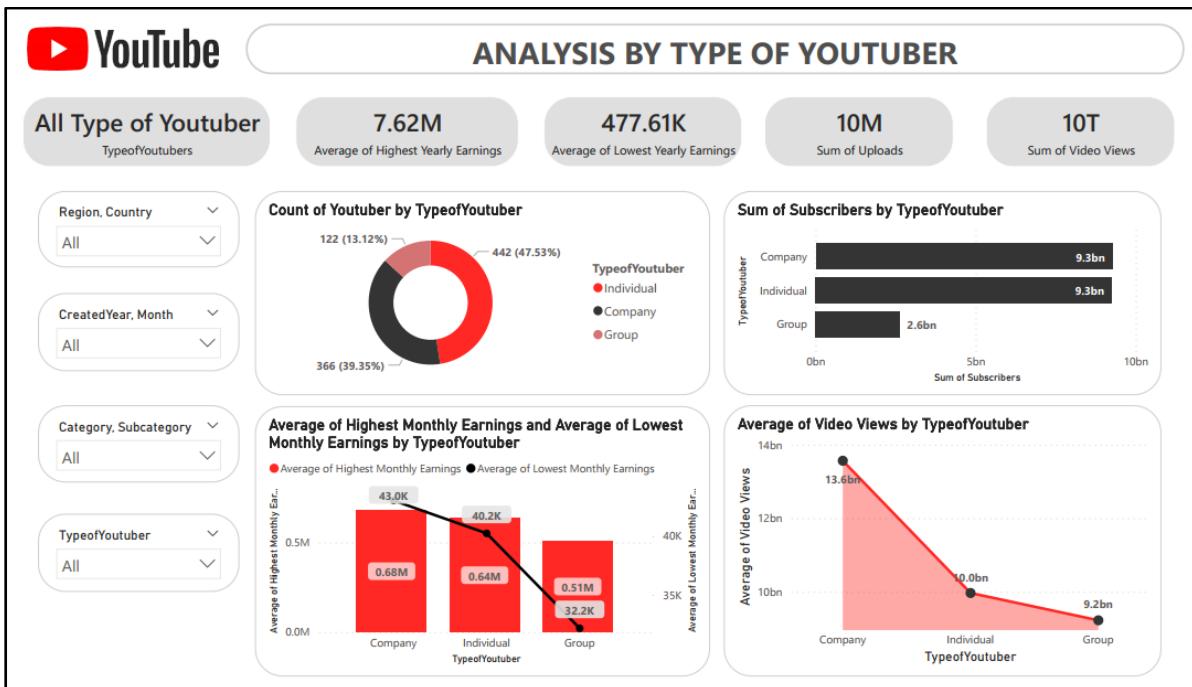
các youtuber kiếm được từ thể loại này có thể lên đến 740 nghìn đô la hàng tháng và ít nhất mỗi tháng là 47.33 nghìn đô la. Xếp sau đó là các chủ đề về Entertainment, Music và Gaming. Đặc biệt, News and Politics mặc dù không nằm trong top 3 chủ đề phổ biến được các youtuber lựa chọn nhưng lại là chủ đề xếp thứ 3 về trung bình thu nhập hàng tháng cao nhất và thấp nhất.

Top 3 chủ đề có lượt views trung bình cao nhất đều liên quan đến giải trí, bao gồm Shows, Trailers và Music. Đặc biệt, số lượt views trung bình của Shows có sự chênh lệch đáng kể so với các chủ đề khác với 42.41 tỷ view. Điều này cho thấy đa phần người xem đang có xu hướng quan tâm đến các chủ đề về giải trí trên Youtube. Một số chủ đề khác cũng nhận được sự quan tâm lớn từ người xem, cụ thể là chủ đề Education xếp ở vị trí thứ 4, Pets & Animals xếp ở vị trí thứ 6, News & Politics xếp ở vị trí thứ 9 và People & Blogs xếp ở vị trí thứ 10.

Một lần nữa, xu hướng quan tâm đến chủ đề giải trí của người xem được thể hiện thông qua số lượt subscribe trung bình. Trong top 5 chủ đề có số lượt subscribe trung bình cao nhất thì đã có 3 chủ đề liên quan đến giải trí, bao gồm Shows, Trailers, Sports với lần lượt 49.37 triệu, 39 triệu và 26.19 triệu lượt subscribe. Trong đó, Shows và Trailers có lượng subscribe cao nhất nhì. Theo sau đó là các chủ đề Nonprofits & Activism xếp ở vị trí thứ 3 và Education xếp thứ 4 với số lượng subscribe trung bình lần lượt là khoảng 27.75 triệu, 26.62 triệu.

Kết luận: Không thể phủ nhận rằng các chủ đề liên quan đến giải trí đang nhận được sự quan tâm sâu sắc từ người xem trong cộng đồng youtube. Tuy nhiên, các chủ đề về Education cũng nhận được sự quan tâm đáng kể, được thể hiện qua số lượt views và subscribes trung bình. Đặc biệt, People & Blogs là chủ đề tuy chỉ xếp thứ 10 về số lượt views trung bình nhưng lại là chủ đề dẫn đầu về mức trung bình thu nhập hàng tháng cao nhất và thấp nhất cho các youtuber.

4.9. Phân tích hiệu suất của các kênh Youtube dựa theo hình thức kênh youtube (Analysis by TypeofYoutuber)



Hình 99 - Phân tích hiệu suất của các kênh Youtube theo TypeofYoutuber

Nhận thấy rất nhiều lợi thế khi hoạt động cá nhân hoặc liên kết với các công ty, những YouTuber thường lập kênh youtube chủ yếu đi theo hai hướng: Hoạt động cá nhân hoặc kết nối với các công ty để vạch ra sẵn con đường phát triển cho nội dung kênh hướng tới. Các kênh YouTube hoạt động cá nhân chiếm tỷ trọng lớn nhất trong các hình thức hoạt động mà các nhà sáng tạo nội dung lựa với hơn 442 kênh (47,53%). Theo sau đó là các kênh theo các kênh có công ty quản lý hỗ trợ với 366 kênh chiếm 39,35% và hình thức hoạt động theo nhóm với 122 kênh khá lép vé so với 2 hình thức còn lại khoảng 13,12%. Lượng người đăng ký (subscribers) của các cá nhân với 9,25 tỷ lượt cũng không hề kém cạnh so với các kênh có những công ty đầu tư sản xuất (khoảng 9,27 tỷ lượt).

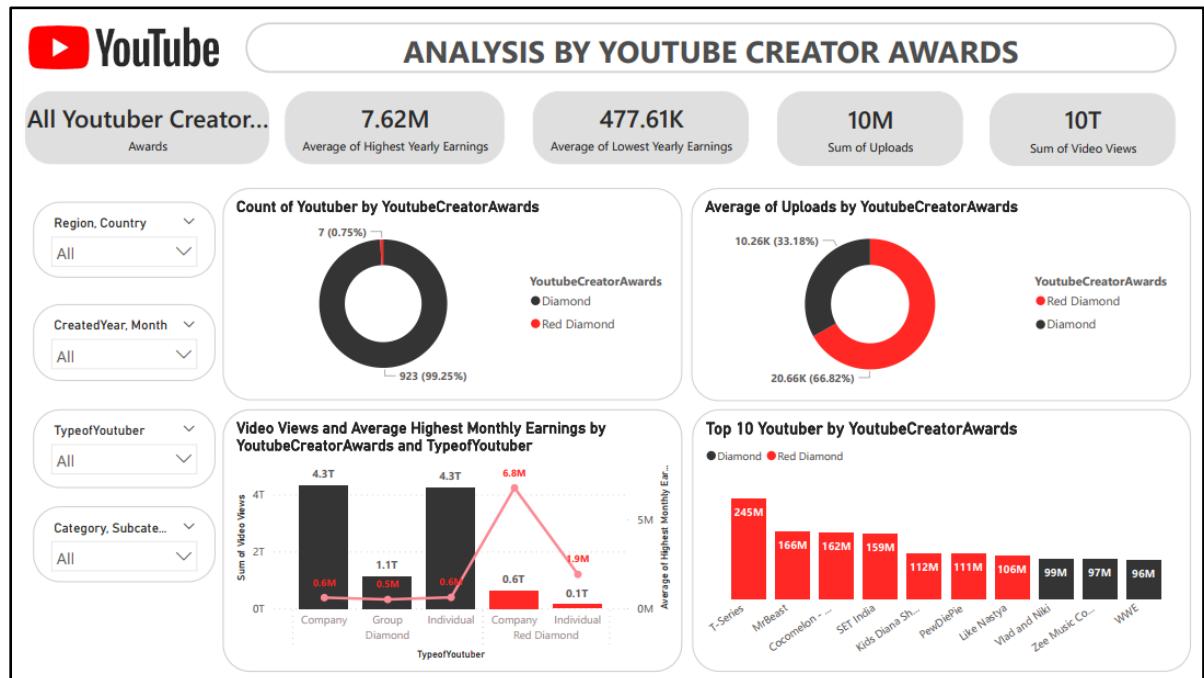
Các kênh cá nhân trên YouTube cho phép người tạo nội dung có sự kiểm soát về nội dung. Họ có thể tự do quyết định về các vấn đề như cách trình bày, ngữ cảnh, ý kiến và phong cách biểu đạt nội dung để thể hiện cá tính và sáng tạo của mình một cách dễ dàng và thu hút những khán giả yêu thích phong cách cá nhân của riêng YouTuber đó. Những kênh có công ty chủ quản thì đã được lên lộ trình sẵn nội dung video, nội dung video và thời gian lên hình đã được chuẩn bị kỹ càng, thường là về các vấn đề “hot” trong xã hội để thu hút người đăng ký quan tâm những chủ đề trên. Các kênh hoạt động nhóm do không có nhiều video đăng tải nên cũng không thu hút

người dùng theo dõi như hai hướng trên.

Không chỉ vượt xa về số lượng youtuber và số lượng subscriber, các kênh hoạt động theo hình thức cá nhân hoặc có các công ty hỗ trợ còn có dẫn đầu mức trung bình thu nhập hàng tháng cao nhất và thấp nhất. Trung bình thu nhập mà các youtuber cá nhân có thể lên đến 0,68 triệu đô la hàng tháng và ít nhất mỗi tháng khoảng 43 nghìn đô la. Xếp theo sau đó là các kênh có công ty chủ quản với khoảng cao nhất là 0,64 triệu đô. Mặc dù lượng subscriber bị bỏ xa nhưng trung bình thu nhập hàng tháng của các kênh hoạt động nhóm lên tới 0,5 triệu đô la và thấp nhất là khoảng 32,2 nghìn đô cho thấy người hâm mộ vô cùng chịu chi để ủng hộ những sản phẩm mà họ quảng cáo hoặc làm đại diện và tiềm năng tăng trưởng của các kênh hoạt động theo nhóm là vô cùng khả quan.

Các kênh youtube hoạt động với công ty chủ quản có lượng view trung bình cao nhất với 13,6 tỷ lượt xem, nhiều hơn hẳn so với các kênh hoạt động cá nhân hoặc theo nhóm lần lượt là 10 tỷ và 9,2 tỷ lượt xem. Lý do là vì các công ty đã lên kế hoạch sẵn cho những gì đã, đang và sẽ đăng tải lên, những video cũng đã được lên lộ trình sẵn về nội dung video và thời gian lên hình đã được chuẩn bị kỹ càng, thường là về các vấn đề “hot” trong xã hội để thu hút lượng khán giả quan tâm những chủ đề trên.

4.10. Phân tích hiệu suất của các kênh Youtube dựa theo hình thức kênh youtube (Analysis by YoutubeCreatorAwards)



Hình 100 - Phân tích hiệu suất của các kênh Youtube theo YoutubeCreatorAwards

Nổi bật với số lượng subscribers cao nhất lần lượt là T-series, MrBeast, Cocomelon... - đều là những kênh Red Diamond.

Bên cạnh đó, tỷ trọng Uploads video trung bình của các kênh Red Diamond cũng gấp đôi (66,82%) so với các kênh Diamond (33,18%). Tuy vậy, trong bộ dataset thu thập được thì chỉ có 7 kênh Red Diamond còn lại hơn 900 kênh (99,2% tổng số kênh) đều là nút Diamond. Một trong những nguyên do được đề cập tới là khoảng cách về tiêu chí đạt được nút Red Diamond so với Diamond có cách biệt rất lớn (Diamond cần 10 triệu lượt đăng ký trong khi Red Diamond là 100 triệu lượt đăng ký) nên cần rất nhiều thời gian và công sức để các kênh có thể gia tăng lượng subscribers. Chính vì có lợi thế sẵn có với lượng subscribers khổng lồ, lượng người theo dõi video của các kênh Red Diamond đều rất cao và doanh thu mang lại cũng là rất lớn. Tuy nhiên, số lượng kênh Red Diamond lại chỉ chiếm tỷ trọng nhỏ và đa số đều là các kênh Diamond. Vì vậy, tổng lượng người xem của hơn 900 kênh có lượt đăng ký dưới 100 triệu lớn hơn so với những kênh Red Diamond. Thu nhập trung bình hàng tháng của các kênh Red Diamond dẫn đầu với 6,8 triệu đô la tập trung vào những kênh có công ty quản lý là chính, với lượng người đăng ký khủng và tần suất video ổn định, ngoài thu nhập từ lượng người theo dõi do youtube trả, các công ty đã liên kết với các kênh để gắn quảng cáo trên video để thu được doanh thu khủng. Theo sau đó vẫn là những kênh cá nhân đạt nút Red Diamond với 1,9 triệu đô la. Các kênh Diamond dù có lượng người xem khủng của hơn 900 kênh những trung bình thu nhập hàng tháng cao nhất lại chỉ chạm ngưỡng 0,5-0,6 triệu đô la.

CHƯƠNG 5: ĐỀ XUẤT GIẢI PHÁP

5.1. Hỗ trợ cơ chế gợi ý nội dung

Hỗ trợ cơ chế gợi ý nội dung có nhiều lượt xem và tìm kiếm về các mảng họ hướng tới cho những youtuber mới chưa có kinh nghiệm hoặc lượng khán giả cố định. Theo dõi các xu hướng nổi bật trong mảng hoạt động và tạo nội dung phù hợp bằng việc sử dụng các công cụ như Google Trends hoặc các trang web theo dõi xu hướng để biết được cung cấp thông tin về số lượng tìm kiếm hàng tháng, những gì đang hot và được tìm kiếm nhiều.

Đánh giá:

- **Thời Gian:**

Nhờ lợi thế sẵn có về sự đa dạng của những chủ đề người xem quan tâm và những thuật toán mà Youtube đang sử dụng để đề xuất video cho khán giả hiện

vẫn đang phát huy rất tốt vai trò của mình. Nhóm tin rằng quá trình phát triển thêm một số tính năng hỗ trợ để phân tích xu hướng người xem mặc dù cần nhiều thời gian để nghiên cứu và kiểm nghiệm. Nhưng khi so với khả năng thu hút sự chú ý từ khán giả và tạo ra những cơ hội rộng mở để những Youtuber mới có thể tiếp cận người xem về lâu dài thì đây là được coi là khoản đầu tư tiềm năng. Hiện tại để cạnh tranh với đối thủ mới nổi như TikTok, thay vì bỏ quá nhiều tiền để mời KOL, KOC để quảng cáo cho nền tảng thì Youtube cũng có thể tận dụng nguồn lực sẵn có giúp những nhà sáng tạo nội dung mới gia tăng lượng người theo dõi cũng là gia tăng sự chú ý của công chúng vào Youtube từ đó một lần nữa khẳng định vị thế của ông lớn trong nền tảng chia sẻ video.

- **Chi phí:**

Việc phát triển những công cụ hỗ trợ những nhà sáng tạo nội dung mà Youtube phải đầu tư là rất lớn. Tuy nhiên, nhờ những tính năng mới này có thể thu hút thêm lượng lớn Youtuber tin tưởng vào sự phát triển vững mạnh của nền tảng và lượt người theo dõi mới cho những nhà sáng tạo đó. Youtube có thể tận dụng khả năng đó để phân tích xu hướng quảng cáo có lượng người bấm bỏ qua ít hơn, phân tích quảng cáo được người dùng ưu ái hơn (những quảng cáo có thần tượng của họ đóng vai, những quảng cáo có phần nhạc dễ chịu, những quảng cáo ngắn dưới 15 giây,...) để tối đa hóa lợi nhuận có thể thu được từ việc gắn quảng cáo.

5.2. Tổ chức chương trình đào tạo

Tổ chức các khóa học hoặc hội thảo trực tuyến để đào tạo kỹ năng sản xuất video và quản lý kênh cho các youtuber mới, với các chiến lược SEO. Ví dụ như:

- Tiêu đề video: Nên mô tả chính xác nội dung video, sử dụng từ 65 đến 70 ký tự, bao gồm từ khóa chính.
- Mô tả video: Mô tả nên đầy đủ, từ 250 đến 700 từ, bao gồm từ khóa trong đoạn đầu và lặp lại 2-4 lần trong toàn bộ mô tả.
- Thẻ (Tags): Sử dụng thẻ để giúp người dùng và công cụ tìm kiếm hiểu nội dung video, kết hợp thẻ tổng quát và cụ thể, khuyến nghị sử dụng từ 10 đến 20 thẻ.
- Thumbnail: Tạo thumbnail tùy chỉnh với độ phân giải cao và hình ảnh hấp dẫn để khuyến khích người dùng nhấp vào video.

Đánh giá:

- **Thời gian:**

- + Có thể triển khai nhanh chóng (chỉ cần chuẩn bị tài liệu và đội ngũ giảng viên). Các lớp học có thể học online, nội dung với số lượng buổi học có thể phù hợp với từng phân khúc học viên. Do đó học viên có thể linh hoạt tham gia theo thời gian biểu của bản thân.
- + Tuy nhiên hiệu quả tiếp thu của học viên không chắc chắn được vì đây là khoá online và còn tuỳ vào khả năng tự học của học viên.

- **Chi phí:**

- + Chi phí ban đầu tương đối thấp, chủ yếu cho việc chuẩn bị tài liệu.
- + Chất lượng chương trình phụ thuộc vào đội ngũ giảng viên. Cùng với đó là khó khăn trong việc quản lý và theo dõi tiến độ của học viên. Có thể lãng phí tài nguyên nếu học viên không quan tâm đến các lớp này.

5.3. Phát triển thêm AI cho Youtuber Analytics

AI có thể giúp đọc số liệu và đưa ra các phân tích như là phân tích xu hướng, dự đoán hiệu suất, tối ưu hoá SEO, phân loại khán giả từ đó youtuber có thể lựa chọn nội dung sáng tạo phù hợp cho từng phân khúc khán giả... Theo như phân tích ở chương 4, các thể loại kênh như Education, People & Blogs là những thể loại kênh phổ biến, không chỉ nhận được sự quan tâm lớn từ người xem mà còn ít cạnh tranh hơn so với chủ đề về giải trí. Đặc biệt, People & Blogs là chủ đề dẫn đầu về mức trung bình thu nhập hàng tháng cao nhất. Bằng việc cung cấp về các xu hướng hiện tại có thể giúp Youtuber không chỉ có nhiều kiến thức về số liệu mà còn hiểu rõ hơn về tiềm năng của các loại kênh và tập trung những content phù hợp cho các tệp khán giả.

Đánh giá:

- **Thời gian:**

- + Google đã có sẵn chatbox Gemini trên Youtube có lợi thế là có nền tảng vững chắc để phát triển chatbox dành riêng cho mình.
- + Tiết kiệm thời gian cho Youtuber trong việc tự phân tích dữ liệu và đưa ra chiến lược. Cùng với đó là có thể cung cấp thông tin chi tiết và chính xác hơn về hiệu suất video và xu hướng người xem.
- + Tuy nhiên cần phải có thời gian để phát triển và hoàn thiện hệ thống AI riêng cho Youtube. Bên cạnh đó việc áp dụng AI có thể gặp một số trở ngại về mặt kỹ thuật.

- **Chi phí:**

- + Tăng hiệu quả kiếm tiền từ kênh Youtube và có thể giúp Youtuber tối ưu hóa chi phí sản xuất video và quảng cáo.
- + Còn về nhược điểm thì tuy có nền Gemini sẵn nhưng để phát triển chatbox riêng cho Youtube cũng cần chi phí thuê đội ngũ kỹ thuật viên chuyên môn cao để vận hành và bảo trì hệ thống.

5.4. Tổ chức chương trình vinh danh (Youtube Awards):

Tổ chức chương trình vinh danh dành cho các youtuber tham gia dưới 2 năm nhưng có thành tích nổi bật. Tùy thuộc vào số lượng youtuber tham gia và tình hình hoạt động của các youtuber mà tiêu chí trao giải sẽ khác nhau. Tuy nhiên, các thành tích nổi bật sẽ bao gồm:

- Youtuber có số lượng video nằm trong top video thịnh hành.
- Kênh youtube có số lượng người đăng ký tăng nhanh trong khoảng thời gian ngắn.
- Video có tỷ lệ tương tác cao từ người xem, bao gồm lượt thích, bình luận, chia sẻ.

Bên cạnh giải thưởng là hiện kim, chương trình còn giúp các youtuber mới nổi tăng độ nhận diện với khán giả và các nhãn hàng. Điều này không chỉ giúp họ tăng sự tương tác với người xem mà còn đem lại những hợp đồng hợp tác từ các nhãn hàng. Từ đó có thể tạo ra nguồn thu nhập mới cho các youtuber.

Đánh giá:

- Thời gian:

Việc tổ chức một chương trình vinh danh chuyên nghiệp đòi hỏi thời gian chuẩn bị nhất định, bao gồm: lên ý tưởng, kêu gọi thí sinh, đánh giá hồ sơ, tổ chức sự kiện,... Việc này có thể mất từ vài tháng đến một năm tùy thuộc vào quy mô của sự kiện và cơ sở hạ tầng tổ chức. Tuy nhiên, một chương trình chuyên nghiệp do Youtube tổ chức có thể thu hút sự chú ý của cộng đồng youtuber và đông đảo khán giả trong thời gian ngắn.

- Chi phí:

Youtube cần cân nhắc cho nhiều loại chi phí như tổ chức sự kiện (thuê địa điểm, trang trí, âm nhạc, ánh sáng, thiết bị phục vụ, cũng như chi phí cho các diễn giả, vũ công, nghệ sĩ biểu diễn), chi phí của giải thưởng và chi phí quảng cáo, tiếp thị. Tuy nhiên, việc hợp tác với các nhãn hàng nổi tiếng và đối tác tiềm năng có thể giúp Youtube được hỗ trợ đáng kể về mặt chi phí và truyền thông. Bên

cạnh đó, Youtube cũng nhận lại được một nguồn doanh thu lớn từ việc bán vé cho khán giả đi xem chương trình.

5.5. Cung cấp nội dung độc quyền

YouTube có thể hợp tác với các nhà sáng tạo nội dung để cung cấp nội dung độc quyền chỉ có trên YouTube mà người dùng không thể tìm thấy trên các nền tảng khác, giúp thu hút người xem đăng ký các kênh. Có thể bao gồm các chương trình truyền hình gốc, phim, video âm nhạc và nội dung do người sáng tạo nổi tiếng sản xuất. Đồng thời việc tạo hỗ trợ những nhà sản xuất nội dung tạo ra nội dung địa phương phù hợp với sở thích và văn hóa của người dùng ở các thị trường khác nhau cũng nên được cân nhắc. Điều này giúp giữ chân và thu hút người dùng trong bối cảnh YouTube đang phải đối mặt với sự cạnh tranh ngày càng tăng từ các nền tảng chia sẻ video khác.

Cách thức hợp tác với YouTube để cung cấp nội dung độc quyền:

- Tham gia YouTube Partner Program (YPP): Đây là chương trình dành cho các nhà sáng tạo đáp ứng các tiêu chí nhất định, cho phép họ kiếm tiền từ nội dung của mình. Khi tham gia YPP, nhà sáng tạo có thể tham gia vào các chương trình cung cấp nội dung độc quyền của YouTube, chẳng hạn như YouTube Originals.
- Hợp tác với YouTube MCN: MCN (Multi-Channel Network) là các mạng lưới hỗ trợ nhà sáng tạo trong việc sản xuất, quản lý và phân phối nội dung. Một số MCN có mối quan hệ hợp tác chặt chẽ với YouTube và có thể giúp nhà sáng tạo tiếp cận các chương trình cung cấp nội dung độc quyền.
- Tự tạo nội dung độc quyền: Nhà sáng tạo có thể tự sản xuất nội dung độc đáo và tải lên kênh YouTube của họ. YouTube có nhiều công cụ và tính năng hỗ trợ nhà sáng tạo trong việc tạo và quản lý nội dung.

Ví dụ: PewDiePie - Kênh YouTube nổi tiếng nhất thế giới với hơn 111 triệu người đăng ký, PewDiePie cung cấp các video game thủ, hài kịch và bình luận. Một số video của PewDiePie là nội dung độc quyền do YouTube Originals sản xuất.

Đánh giá:

- **Thời gian** sản xuất nội dung độc quyền trên YouTube có thể thay đổi đáng kể tùy thuộc vào nhiều yếu tố, bao gồm: loại nội dung, độ dài nội dung, chất lượng sản xuất, kỹ năng và kinh nghiệm của nhà sáng tạo nội dung.

Vd:

- + Loại nội dung:

- Video: Quay phim, dựng phim, chỉnh sửa có thể mất từ vài giờ đến vài tuần, tùy thuộc vào độ phức tạp của nội dung.
 - Livestream: Livestream có thể diễn ra trong vài phút hoặc vài giờ, nhưng cần thời gian chuẩn bị trước và sau khi phát sóng.
 - Podcast: Ghi âm, chỉnh sửa podcast có thể mất từ vài giờ đến vài ngày.
- + Bài viết: Viết bài có thể mất từ vài phút đến vài giờ, tùy thuộc vào độ dài và nội dung bài viết.
- + Độ dài nội dung: Nội dung dài hơn thường đòi hỏi nhiều thời gian sản xuất hơn.
- + Chất lượng sản xuất: Nội dung chất lượng cao thường đòi hỏi nhiều thời gian và công sức hơn để sản xuất.
- + Kinh nghiệm và kỹ năng: Nhà sáng tạo có kinh nghiệm và kỹ năng có thể sản xuất nội dung nhanh hơn so với người mới bắt đầu.
- Chi phí sản xuất nội dung độc quyền trên YouTube cũng có thể thay đổi đáng kể tùy thuộc vào nhiều yếu tố, bao gồm: loại nội dung, độ dài nội dung, chất lượng sản xuất.
 - + Loại nội dung:
 - Video: Chi phí cho thiết bị quay phim, phần mềm chỉnh sửa, diễn viên, địa điểm,...
 - Livestream: Chi phí cho thiết bị phát sóng, phần mềm livestream, địa điểm,...
 - Podcast: Chi phí cho thiết bị ghi âm, phần mềm chỉnh sửa, studio,...
 - Bài viết: Chi phí cho phần mềm viết bài, hình ảnh, video,...
 - + Chất lượng sản xuất: Nội dung chất lượng cao thường tốn kém hơn để sản xuất do cần sử dụng thiết bị, phần mềm và dịch vụ tốt hơn.
 - + Địa điểm quay: Quay phim ở những địa điểm đắt đỏ có thể làm tăng chi phí sản xuất.
 - + Bản quyền âm nhạc: Sử dụng nhạc có bản quyền trong video có thể tốn chi phí.

5.6. Tận dụng Youtube Shorts

Như đã đề cập ở chương 2, Youtube Shorts hiện đang trở thành một hiện tượng với lượng truy cập rất lớn. Đây là một cơ hội tốt để tận dụng và phát triển nội dung

ngắn, hấp dẫn để thu hút khán giả. Bởi vì Youtube Shorts có thời lượng chỉ từ 15s đến 1 phút nên có thể giúp truyền đạt thông điệp một cách nhanh chóng và dễ dàng hơn cho người xem. Từ đó thì có thể giúp tạo điều kiện cho việc tương tác mạnh mẽ và thúc đẩy số lượng người xem tăng nhanh. Cụ thể, các nhà sáng tạo nội dung có thể chuyển đổi các video dài trên Youtube thành các video ngắn hơn dưới dạng Youtube Shorts với nội dung quan trọng vẫn được làm nổi bật. Và Youtube Shorts có thể dễ xuất hiện ở phần đề xuất hơn là các video dài. Điều này không chỉ tận dụng tối đa nội dung video gốc mà còn tạo ra nhiều cơ hội để thu hút thêm người xem mới. Với sự phát triển mạnh mẽ của nền tảng Youtube Shorts thì việc đầu tư thêm vào ý tưởng tạo ra nội dung ngắn và hấp dẫn là một bước đi có thể mang lại nhiều lợi ích cho các nhà sáng tạo nội dung.

Đánh giá:

- Thời gian:**

Việc tạo nội dung ngắn trên Youtube tiết kiệm thời gian hơn so với việc phát triển các video dài. Nhờ thời lượng ngắn, các nhà sáng tạo nội dung tập trung nhanh chóng vào ý tưởng chính và thực hiện nó một cách nhanh chóng hơn. Quá trình chuyển đổi video dài thành một Shorts có thể mất một ít thời gian ban đầu để chỉnh sửa và cắt ghép, nhưng sau khi đã quen với công việc này thì quy trình này sẽ trở nên đơn giản hơn. Việc tận dụng Youtube Shorts không chỉ giúp tối ưu hóa thời gian sản xuất mà còn tạo ra nhiều cơ hội để thu hút người xem mới.

- Chi phí:**

Việc tạo một video Shorts không đòi hỏi chi phí lớn. Các nhà sáng tạo nội dung chỉ cần một thiết bị quay phim chất lượng và các công cụ edit video cơ bản. So với việc sản xuất các video dài, việc tạo Shorts có thể giúp tiết kiệm đáng kể thời gian và công sức. Vậy nên, Youtube Shorts mang lại một cơ hội tuyệt vời để tạo ra nội dung hấp dẫn mà không cần phải bỏ ra quá nhiều chi phí.

5.7. Đa dạng hóa phương thức kiếm tiền cho nhà sáng tạo nội dung

YouTube có thể cung cấp hỗ trợ tài chính cho các nhà sáng tạo nội dung để giúp họ tạo ra nội dung chất lượng cao hơn. Bên cạnh đó, nền tảng nền sử dụng mô hình chia sẻ doanh thu linh hoạt hơn, thay vì sử dụng mô hình CPM (chi phí cho mỗi nghìn lượt xem) cố định, có thể sử dụng mô hình chia sẻ doanh thu linh hoạt hơn, chẳng hạn như mô hình chia sẻ doanh thu theo tỷ lệ phần trăm hoặc mô hình dựa trên hiệu suất. Cung cấp nhiều tùy chọn kiếm tiền hơn cũng là một chính sách cần thiết.

YouTube có thể cung cấp nhiều tùy chọn kiếm tiền hơn cho người sáng tạo nội dung, chẳng hạn như cho phép họ bán hàng hóa và dịch vụ trực tiếp trên nền tảng hoặc thông qua các kênh khác như Patreon. Bằng cách thực hiện những thay đổi này, YouTube có thể tạo ra một môi trường bền vững hơn cho người sáng tạo nội dung và đảm bảo rằng họ được trả công một cách công bằng cho công việc của họ.

Đánh giá:

- **Thời gian:**

Hỗ trợ tài chính cho nhà sáng tạo nội dung có thể được thực hiện ngay lập tức, nhưng cần thời gian để xem xét các yêu cầu và quyết định mức độ hỗ trợ từ bộ phận quản trị tài chính của Youtube. Về thay đổi mô hình chia sẻ doanh thu, việc này có thể mất thời gian để phát triển và thử nghiệm các mô hình mới và đánh giá hiệu quả của chúng, cũng như để nhận phản hồi từ cộng đồng người sáng tạo.

- **Chi phí:**

Chi phí cho việc hỗ trợ tài chính có thể rất cao, tùy thuộc vào số lượng và mức độ hỗ trợ cho mỗi người sáng tạo. Cần xây dựng bộ tiêu chí đánh giá, tiêu chuẩn đối với các kênh đủ điều kiện để được nhận hỗ trợ tài chính. Thay đổi mô hình chia sẻ doanh thu bao gồm chi phí phát triển hệ thống mới, cũng như chi phí tiếp thị và giáo dục để người sáng tạo hiểu về mô hình mới.

CHƯƠNG 6: KẾT LUẬN

6.1. Kết luận

Sau quá trình thực hiện bài đồ án phân tích tìm hiểu về Youtube thông qua bộ dataset Youtube Global Statistics, nhóm đã hoàn thành việc phân tích về tình hình doanh nghiệp và đề xuất những xu hướng phát triển hợp lý để giúp các Youtubers nói chung có thể dễ dàng đưa ra phương hướng giúp kênh duy trì tương tác với người dùng cũ và gia tăng subscribers mới. Đồng thời, nhóm cũng đã đề xuất một số phương pháp về hệ thống vận hành của Youtube nói riêng để gia tăng sức cạnh tranh và duy trì vị thế đầu bảng trong những nền tảng chia sẻ video trực tuyến.

Từ các thuộc tính có sẵn, nhóm đã thực hiện tiền xử lý dữ liệu bằng nhiều phương pháp khác nhau như đã trình bày ở chương 3 sau đó phân tích để phân ra các bảng dim, fact và đưa ra dạng lược đồ hình ngôi sao để tiến hành phân tích sâu bằng Power BI. Mặc dù đã rất nỗ lực tuy nhiên nhóm cũng gặp nhiều vấn đề trong quá trình

phân tích các thuộc tính để tạo các dim và cần có gắng hơn nữa để cải thiện hiệu quả phân tích, đánh giá doanh nghiệp từ nhiều khía cạnh khác nhau từ bộ dữ liệu.

6.2. Hướng phát triển

Sau quá trình phân tích dữ liệu trên, nhóm đã nắm bắt được các kiến thức cần thiết để phân tích và đưa ra các đề xuất phát triển cho doanh nghiệp và Youtubers. Vậy nên, nhóm hoàn toàn có khả năng sử dụng một bộ dữ liệu của một công ty thực tế để phân tích tình hình kinh doanh và đưa ra các đề xuất về chiến lược kinh doanh giúp công ty đưa ra các quyết định nhằm nâng cao hiệu suất kinh doanh và hoạt động của doanh nghiệp một cách thuận lợi hơn trong tương lai. Ngoài ra, đồ án này của nhóm có thể được sử dụng như một nguồn tài liệu tham khảo cho những báo cáo có chủ đề hoặc hướng đi tương tự. Hơn nữa, đồ án phân tích của nhóm còn có tiềm năng để thể phát triển và hoàn thiện hơn ở tương lai, thông qua việc phân tích sâu hơn và thực hiện các giải pháp chiến lược đã đề xuất ở chương 5 một cách quan hơn để áp dụng vào thực tiễn doanh nghiệp.

TÀI LIỆU THAM KHẢO

1. Brian Dean, 2024, *How Many People Use YouTube? [New Data]*, viewed 20 April 2024, <https://backlinko.com/youtube-users#most-used-social-media-platforms>
2. ThS. Phạm Thị Thanh Tâm, 2024, Slide bài giảng bộ môn Phân tích nghiệp vụ kinh doanh.