

ANALYZING CONTENT CREATION AND CONSUMPTION ON YOUTUBE THROUGH VISUAL ANALYTICS.

Prof. Nikita Ramrakhiani
Assistant Professor- Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
nikita.ramrakhiani@ves.ac.in

Ms. Hiya Bhatia
Student- Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
hiya.bhatia2224p@ves.ac.in

Ms. Preety Gind
Student- Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
preety.gind2224p@ves.ac.in

Mr. Ajinkya Kedare
Student- Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
ajinkya.kedare2224p@ves.ac.in

Abstract—In today's digital landscape, social media platforms hold a pivotal role in disseminating content, nurturing interactions, and shaping user preferences. This comprehensive exploration delves into prominent YouTube channels, assessing subscribers, video views, and upload frequency across diverse categories and countries. As of January 2023, YouTube stands as a global content powerhouse with 2.5 billion monthly and 1.2 billion daily active users, engaging with a billion hours of content daily. With a rapid influx of 500 hours of video uploaded each minute, it remains an unparalleled hub for content creation. The platform's impact is global, notably in countries like India, the US, and Indonesia, with languages like English, Hindi, and Spanish reigning supreme. Categories spanning music, entertainment, gaming, education, and news captivate users, who dedicate an average of 2 hours and 24 minutes per session, viewing around 100 videos per month. The financial clout of YouTube is underscored by its \$28.8 billion revenue in 2022. The integration of sociodemographic factors provides a broader comprehension of the intricate interplay between content engagement and societal attributes. Revealing captivating patterns and trends, this analysis unveils the evolving terrain of YouTube's content creation and audience engagement. Diverse data visualizations, including multi attribute plots to depict channel distribution, popularity, and interaction dynamics have been used in this research. This amalgamation extends beyond content creators and marketers, offering researchers valuable insights into digital platform dynamics within the broader context of societal indicators and evolving social media trends. Temporal trends in video views and subscribers are effectively conveyed through line charts, illuminating growth trajectories. This analysis delves into social media trends, acknowledging their symbiotic rapport with YouTube. Crucially, this research aimed at finding how analytics reflects in the domains of social media and contributes towards refining strategies, optimizing engagement, and honing targeting efforts and acquiring ways to maximize the synergy that magnifies reach and cultivates cross-platform expansion.

Keywords—(Social Media Analytics , Click-through rate, Data Visualisation using Tableau, Machine Learning, Predictive Modeling.)

I. INTRODUCTION

As one of the largest social platforms, YouTube holds a prominent position as both a dominant video-sharing platform and a widely recognized search engine. Users turn to YouTube to access a diverse range of content, including instructional videos, lectures, topical discussions, and skill development resources. With a global user base, YouTube stands as the foremost online video destination, witnessing an astonishing 4 billion hours of video consumption each month, accompanied by the continuous upload of 72 hours of new video content per minute.

This study introduces a comprehensive data analytics workflow tailored specifically for the analysis of YouTube data. This analytics framework encompasses a wide spectrum of metrics that revolve around channel performance, encompassing factors such as video views, likes, dislikes, comments, and the number of subscribers, among others.

Within the technical landscape, it is crucial to acknowledge the pervasive phenomenon of user impatience, particularly on platforms like YouTube. This behavior, characterized by swift navigation and content skipping, carries significant implications for content creators, marketers, and businesses leveraging YouTube for engagement and promotional purposes. Effectively addressing this behavior necessitates the strategic design of content and the optimization of viewer interactions to ensure viewer retention and engagement.

Furthermore, maintaining a consistent and captivating narrative throughout the video's duration is paramount to sustain and nurture viewer interest and involvement. In essence, the prevalence of viewer impatience on YouTube underscores the importance of crafting concise, impactful, and pertinent content. It serves as a reminder that, in the digital era, where distractions are abundant, content that captivates and retains attention is most likely to thrive and convey the intended message or entertainment value.

Our research interests are succinctly encapsulated in the following points:

1. Identifying the category that garners the highest number of video views.
2. In the pursuit of maximizing earnings, determining whether it is the quantity of video views or the volume of subscribers that exerts the most substantial influence on an influencer's income.

II. RESEARCH METHODOLOGY

Visual analytics is an interdisciplinary field that converges data analysis, statistical methodologies, and advanced data visualization techniques into a cohesive framework. Its primary aim is to facilitate the comprehension and extraction of insights from intricate and voluminous datasets through the deployment of interactive visual interfaces and analytical tools.

Visualization tools can be employed to visualize distributions and conduct statistical tests. Box plots and violin plots, for instance, are valuable for comparing data distributions and assessing statistical significance.

Visualization tools are fundamental for EDA, allowing researchers to generate and validate hypotheses, identify potential variables of interest, and iteratively refine research questions.

visualization tools are indispensable assets in analytics research, as they empower researchers to efficiently explore, understand, and communicate complex data. These tools facilitate the extraction of actionable insights, enable data-driven decision-making, and enhance the overall rigor and efficacy of research endeavors.

Logarithmic transformations are useful for compressing data with exponential growth, making it more linear and easier to analyze. They are often applied when data exhibits a wide range of values, and you want to reduce the impact of extreme values on the analysis.

Square root transformations are employed to reduce the impact of outliers and make data distributions more symmetric. They are beneficial when dealing with count data or data with a right-skewed distribution.

Exponential transformations are used to emphasize exponential growth or to linearize data with exponential decay. They can be applied when you want to model data with a rapid rate of increase or decrease.

These transformations serve various purposes in data analysis, such as normalizing data distributions, reducing the impact of outliers, and making relationships between variables more linear. The choice of transformation depends on the specific characteristics of the data and the objectives of the analysis. Data analysts and statisticians often apply these transformations to prepare data for statistical modeling, hypothesis testing, and further analysis. convert into technical language

III. EXPLORATORY DATA ANALYSIS

A. Data Cleaning

Exploratory Data Analysis (EDA) is an essential preliminary step in comprehending a dataset, such as YouTube data statistics. This process involves the collection, cleaning, and analysis of the vast amount of data generated by YouTube, aiming to unveil trends, patterns, and valuable insights. data cleaning and preprocessing, which includes handling missing data points, dealing with outliers, and converting data types to make it suitable for analysis. After these preparatory steps, the exploratory analysis phase can commence. One of the fundamental insights derived from YouTube data pertains to understanding the popularity of different types of videos. By analyzing metrics such as views, likes, and subscribers, patterns in content consumption can be discerned. For instance, EDA might reveal that music videos, educational videos, or tutorials consistently garner higher views and engagement than other content categories. This research underscores the paramount importance of judiciously addressing missing values as a crucial preparatory step in the creation of meaningful data visualizations. The strategic choice to remove all instances of missing values, based on a careful assessment of their prevalence in the dataset, serves as an effective approach to uphold data integrity while simplifying the dataset for the purpose of visualization. By implementing such meticulous data preparation techniques, analysts and researchers ensure that the resulting visualizations are not only informative but also founded on a robust and complete dataset. This research paper delves into the treatment of outliers during the data analysis process. An integral facet of this analysis involves the deliberate choice of not addressing outliers due to their negligible impact on the dataset. The decision to forego outlier treatment is made after conducting a rigorous examination, which ascertains that the presence of outliers does not exert any significant influence on the dataset's overall statistical characteristics.

B. Data Preprocessing

The intricacies surrounding data scaling during the data preprocessing phase of analytics are noteworthy. Despite rigorous endeavors to normalize or standardize the dataset, it became evident that additional transformations were necessary. To address this issue, we employed a logarithmic scaling technique due to the observed non-normality within the dataset. This strategic application of logarithmic scaling was implemented to mitigate the non-normal distribution of the data.

C. Data Transformation

Linear regression is a statistical technique employed in analytical research to establish and quantify the association between a dependent variable (Y) and one or more independent variables (X), with the objective of fitting a linear equation to observed data. This methodology is integral to the field of statistics and machine learning, facilitating the comprehension and prediction of intricate relationships among variables.

Linear regression aims to find the best-fitting linear equation that describes the relationship between the dependent variable (Y) and one or more independent

variables (X). The equation for a simple linear regression with one independent variable can be represented as:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_pX_p+c$$

Y is the dependent variable (the target or response variable). $X_1, X_2, \dots, X_p$ are the independent variables (predictors or features).

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the coefficients (parameters) to be estimated.

c is the error term, representing the unexplained variance or noise in the relationship.

The goal of linear regression is to estimate the coefficients such that the linear equation best fits the observed data points. This is typically done using a method called Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values.

Root Mean Square Error (RMSE) is a widely employed statistical and machine learning metric for assessing the precision of predictive models, particularly those designed for regression tasks. It quantifies the average magnitude of discrepancies between predicted values and actual values within a dataset. RMSE is particularly valuable when evaluating the performance of models that generate continuous predictions, such as linear regression, decision trees, and neural networks.

The Mean Absolute Error (MAE) is a widely employed metric in both statistical analysis and machine learning, primarily utilized for assessing the predictive accuracy of models, particularly within regression tasks. MAE quantifies the average absolute deviation between predicted values and actual values within a given dataset.

Bias in the context of modeling pertains to the discrepancy introduced when approximating a complex real-world problem with a simplified model. A model exhibiting high bias demonstrates limited responsiveness to the training data, resulting in an oversimplified representation of the underlying patterns. Consequently, systematic errors emerge in the model's predictions. In more technical terms, elevated bias signifies underfitting, indicating that the model inadequately captures the intricacies of the data.

Variance, conversely, denotes the error stemming from the model's susceptibility to minor fluctuations within the training data. A model characterized by high variance displays an excessive inclination toward the training data, often capturing noise or minor irregularities.

The resulting bar chart effectively visualized the distribution of video views across countries. The findings revealed distinct patterns of video viewership among different nations. Interpreting the results of this visualization, we can discern that the histogram discover.the distribution of video views across different countries. Out of 42 Countries in the data. From the chart, it is evident that India and the United State (US) exhibit notably higher video view counts compared to other countries. This implies that content related to these two regions has garnered more attention or engagement, possibly due to various factors such as population size, content popularity, or marketing efforts. This insight can be valuable for content creators or marketers seeking to target specific audiences or regions based on their video view preferences.

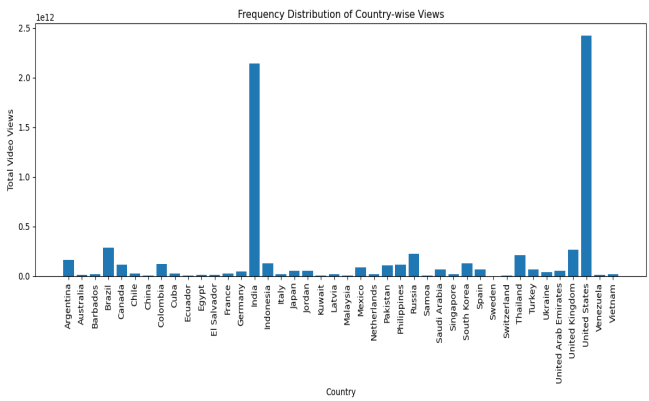


Fig 1. Frequency distribution of Country- wise

The histogram-based analysis of category-wise video views distribution underscores the pronounced disparities in viewership across diverse content categories. While "Music," "Entertainment," and "People and Blogs" reign as the top three categories, "Travel and Events," "NonProfits and Activism," "Autos & Vehicles," and "Movies" emerge as the categories with the least viewership. These insights serve to inform content creators and marketers, facilitating data-driven decisions and strategic content development in the dynamic landscape of online video viewership. These categories are less popular among the audience.

1. Cateogy wise video views

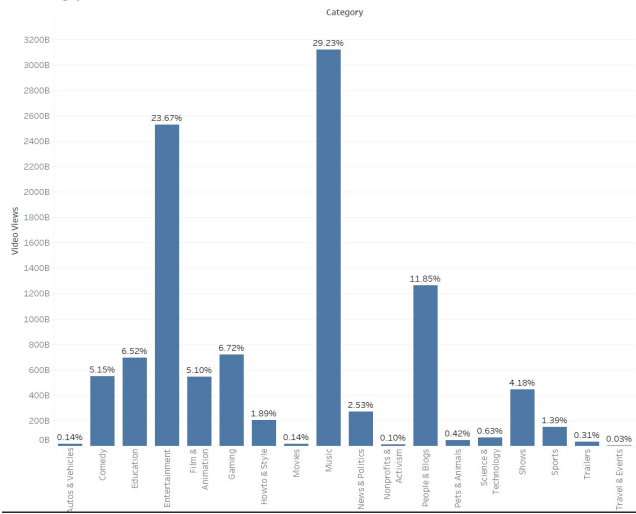


Fig 2. Category wise video views

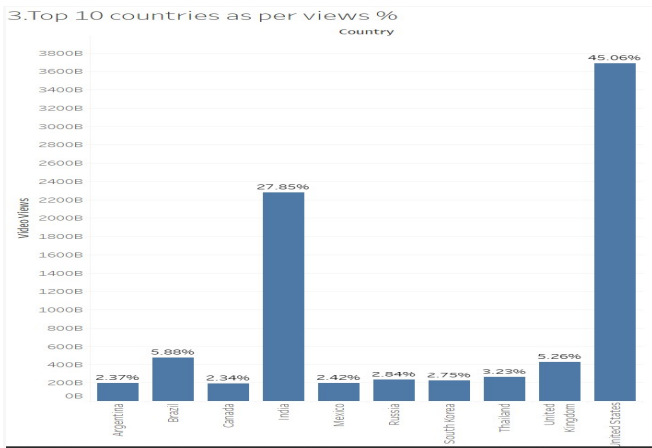


Fig 3. Top 10 Countries as per views

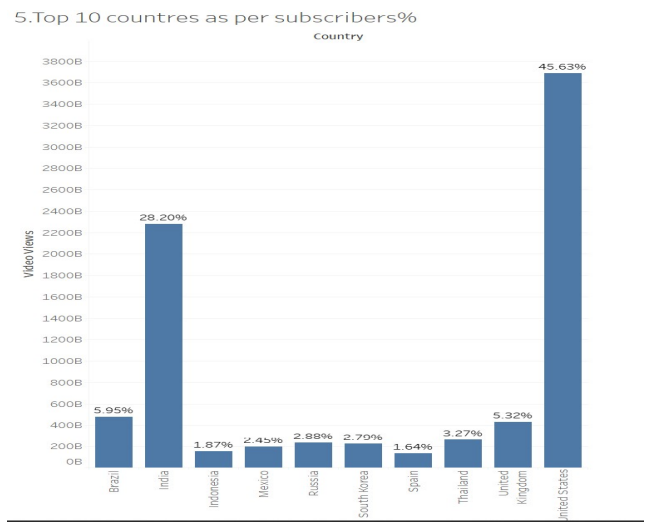


Fig4. Top countries as per subscribers

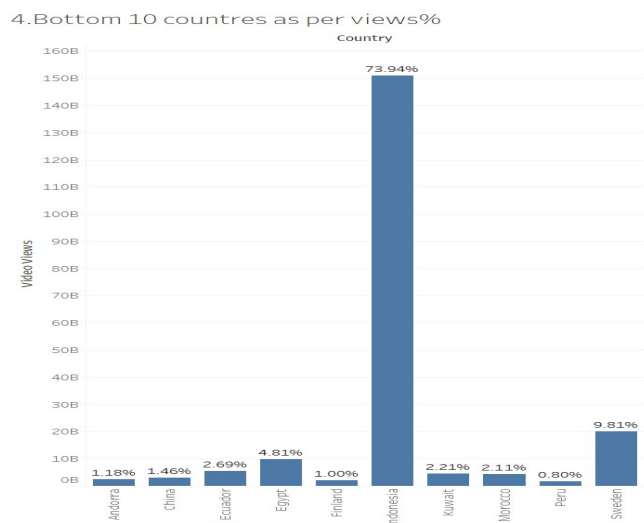


Fig 5. Bottom 10 Countries as per views

At the apex of the viewership hierarchy, the United States emerges as the dominant force, with a commanding share of 45% of total video views. This preeminent position underscores the United States' significance as a primary market for YouTube content consumption. Conversely, at the opposite end of the spectrum, the analysis reveals that Peru holds the lowest viewership share among the sampled countries, accounting for a mere 0.80% of total video views. This diminutive percentage underscores the relatively limited YouTube viewership within the Peruvian market. The divergence in video views between the top-ranking United States and the bottom-ranking Peru epitomizes the substantial variations in YouTube viewership across different countries. Such insights are of paramount importance for content creators, marketers, and platform administrators, as they inform strategic decisions pertaining to content.

The focal point of this section is the identification and quantification of extremities, specifically, the top and bottom countries in terms of subscriber counts. Through rigorous data analysis, this paper sheds light on the countries boasting the highest and lowest percentages of total subscribers, thus offering valuable insights into YouTube's global subscriber landscape. At the zenith of the subscriber hierarchy, the United States emerges as the dominant entity, commanding an impressive 45.63% of the total subscriber base. This prominent position not only underscores the United States' substantial presence within the YouTube ecosystem but also highlights its significance as a primary market for content creators. Conversely, at the opposite end of the spectrum, the analysis reveals that Peru occupies the lowest rung among the sampled countries, contributing a modest 2.41% to the total subscriber count. This relatively modest percentage accentuates the comparatively lower subscriber base within the Peruvian market. The substantial contrast in subscriber counts between the top-ranking United States and the bottom-ranking Peru epitomizes the diverse and heterogeneous nature of YouTube's global subscriber landscape.

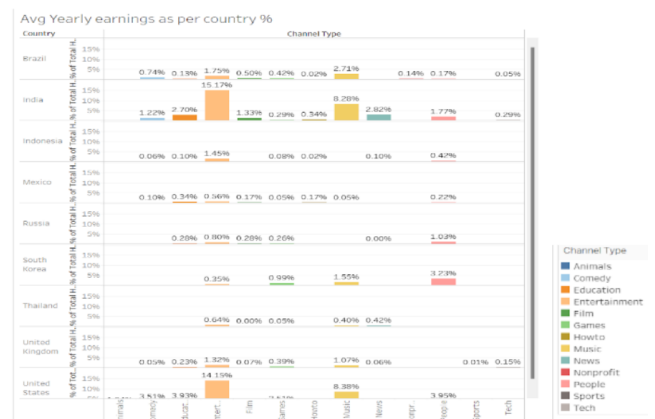


Fig6. Average yearly earnings with categories

The symbiotic relationship between content categories and earnings elucidates the economic underpinnings of YouTube's global ecosystem. The preponderance of "Music" and "Entertainment" as both high-viewed and high-earning categories substantiates their attractiveness to content creators and advertisers alike. This dual appeal not only bolsters viewership but also fosters monetization opportunities. This analysis underscores the intricate interplay between content categories, viewership, and earnings within the YouTube platform. "Music" and "Entertainment" categories, consistently marked by high viewership, also wield significant influence in driving revenue, a phenomenon that resonates across multiple countries.

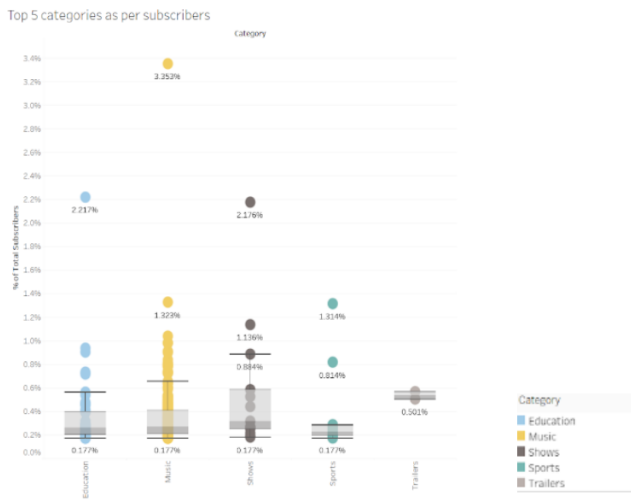


Fig.7 Boxplot of Top 5 categories

The analysis reveals that Education, Music, and Shows are the fastest-growing categories, followed by Sports and Trailers. Content creators should track growth rates over time and consider competitive landscapes to identify the most promising categories.

RESULTS

Implementation of Model

In the data preprocessing phase of analytics, despite diligent normalization and standardization efforts, the presence of non-normality in the dataset necessitated the use of logarithmic scaling as a strategic transformation to address this issue and mitigate the non-normal data distribution.

In this specific iteration of our research within the domain of business analytics, we employed backpropagation as the chosen optimization algorithm. This technique was implemented using R Studio, incorporating all available variables as part of our model input, ensuring comprehensive coverage and analysis of the dataset.

```
lm(formula = Avg_yearly_income ~ rank + video.views + category +
+uploads + video_views_rank + video_views_for_the_last_30_days +
+subscribers + subscribers_for_last_30_days, data = you_3)
```

Residuals:	Min	1Q	Median	3Q	Max
	-31637251	-1201095	-89502	1131659	39786980

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.551e+06	3.134e+06	1.452	0.1470
rank	-1.892e+03	8.272e+02	-2.288	0.0225 *
video_views	3.503e-04	2.279e-05	15.375	< 2e-16 ***
categoryComedy	-8.931e+05	3.138e+06	-0.285	0.7761
categoryEducation	-2.247e+06	3.157e+06	-0.712	0.4769
categoryEntertainment	-1.955e+06	3.086e+06	-0.633	0.5267
categoryFilm & Animation	-2.720e+06	3.173e+06	-0.857	0.3918
categoryGaming	-1.956e+06	3.125e+06	-0.626	0.5316
categoryHowto & Style	-1.387e+06	3.292e+06	-0.421	0.6737
categoryMovies	-2.791e+06	4.337e+06	-0.644	0.5201
categoryMusic	-1.963e+06	3.094e+06	-0.634	0.5261
categorynan	-1.866e+05	3.152e+06	-0.059	0.9528
categoryNews & Politics	-2.105e+06	3.287e+06	-0.640	0.5222
categoryNonprofits & Activism	-1.142e+06	4.376e+06	-0.261	0.7942
categoryPeople & Blogs	-2.001e+06	3.107e+06	-0.644	0.5197
categoryPets & Animals	-1.729e+06	3.954e+06	-0.437	0.6622
categoryScience & Technology	-3.853e+06	3.340e+06	-1.153	0.2492
categoryShows	-1.341e+06	3.325e+06	-0.403	0.6869
categorySports	-4.659e+05	3.390e+06	-0.137	0.8907
categoryTrailers	-3.507e+06	4.344e+06	-0.807	0.4199
uploads	7.349e+00	5.352e+00	1.373	0.1703
video_views_rank	-5.960e-01	2.698e-01	-2.209	0.0276 *
video_views_for_the_last_30_days	5.156e-03	4.327e-04	11.917	< 2e-16 ***
subscribers	-1.891e-01	2.254e-02	-8.390	3.94e-16 ***
subscribers_for_last_30_days	7.075e+00	3.586e-01	19.733	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4331000 on 563 degrees of freedom
Multiple R-squared: 0.7564, Adjusted R-squared: 0.7461
F-statistic: 72.86 on 24 and 563 DF, p-value: < 2.2e-16

While the R-square value in our analysis indicates an acceptable degree of variance explained by the model, it is important to note that the statistical significance levels associated with the variables under consideration were relatively low. The p-value serves as a pivotal statistical measure, playing a crucial role in aiding researchers in evaluating hypothesis within the framework of hypothesis testing. The relationship between the variables in our analysis is deemed insignificant due to the p-value exceeding the threshold of 0.05. This finding states that there is need for further investigation and refinement of the model to better capture the underlying dynamics within the data.

In the first iteration of our analysis, a notable observation emerged: among all the categories, "comedy" exhibited a significantly substantial influence on the average yearly income. However, in the second iteration, the significance of this relationship was reaffirmed through rigorous statistical testing, further reinforcing the notion that the "comedy" category significantly impacts average yearly income. This finding is a valuable contribution to our business analytics research.

```
> model14 <- lm(Avg_yearly_income ~ rank + video.views + uploads+is_comedy +
+ video_views_rank + video_views_for_the_last_30_days +
+ subscribers + subscribers_for_last_30_days,data=you_4)
> summary(model14)
```

Call:

```
lm(formula = Avg_yearly_income ~ rank + video.views + uploads +
+ is_comedy + video_views_rank + video_views_for_the_last_30_days +
+ subscribers + subscribers_for_last_30_days, data = you_4)
```

Residuals:	Min	1Q	Median	3Q	Max
	-31397395	-1052262	-9561	950678	41297062

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.505e+06	6.723e+05	3.726	0.000214 ***
rank	-1.624e+03	8.059e+02	-2.016	0.044298 *
video_views	3.533e-04	2.181e-05	16.201	< 2e-16 ***
uploads	6.856e+00	4.218e+00	1.625	0.104610
is_comedy	9.838e+05	7.121e+05	1.382	0.167646
video_views_rank	-6.148e-01	2.669e-01	-2.303	0.021611 *
video_views_for_the_last_30_days	5.112e-03	4.212e-04	12.137	< 2e-16 ***
subscribers	-1.905e-01	2.196e-02	-8.673	< 2e-16 ***
subscribers_for_last_30_days	7.180e+00	3.459e-01	20.759	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4311000 on 579 degrees of freedom
Multiple R-squared: 0.7518, Adjusted R-squared: 0.7483
F-statistic: 219.2 on 8 and 579 DF, p-value: < 2.2e-16

In our analysis, it was established that the variables "video views ranks" and "ranks" displayed statistical significance

concerning the dependent variable. Consequently, we proceeded with these variables in our model. However, it is important to note that despite their significance, there was no observed improvement in both the R-squared and Adjusted R-squared values. This outcome indicates that the inclusion of these variables did not lead to an enhanced explanatory power of the model in explaining variations in the dependent variable, as evidenced in our business analytics research.

The distribution of residuals was deemed unacceptable in our analysis, primarily due to the presence of non-normality within the data. In response, we implemented data transformation techniques to address this issue. Additionally, we made the decision to exclude variables based on insights from previous iterations of our research. Subsequently, we arrived at the final model, as illustrated in the accompanying figure. This model reflects the culmination of our efforts in the realm of business analytics, where we meticulously refined our approach to address data distribution concerns and arrived at a more robust and optimized model for our research objectives.

```
> model17 <- lm(Avg_yearly_income_LOG ~ rank + video.views
+               + video_views_rank + video_views_for_the_last_30_days +
+               + subscribers + subscribers_for_last_30_days, data=you_5)
> summary(model17)
```

Call:
lm(formula = Avg_yearly_income_LOG ~ rank + video.views + video_views_rank +
video_views_for_the_last_30_days + subscribers + subscribers_for_last_30_days,
data = you_5)

Residuals:

	Min	1Q	Median	3Q	Max
	-11.0298	-0.4624	0.3173	0.7679	7.6562

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.187e+01	2.351e-01	93.055	< 2e-16 ***
rank	-1.737e-03	2.818e-04	-6.162	1.36e-09 ***
video.views	2.850e-11	9.088e-12	3.136	0.0018 **
video_views_rank	-4.509e-06	9.361e-08	-48.168	< 2e-16 ***
video_views_for_the_last_30_days	3.060e-09	3.284e-10	9.318	< 2e-16 ***
subscribers	-3.415e-08	8.074e-09	-4.230	2.73e-05 ***
subscribers_for_last_30_days	4.014e-07	1.584e-07	2.534	0.0115 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.486 on 569 degrees of freedom
Multiple R-squared: 0.8473, Adjusted R-squared: 0.8457
F-statistic: 526.3 on 6 and 569 DF, p-value: < 2.2e-16

In the final model, it was determined that both video views and subscribers significantly contribute to the income of influencers or advertisers. However, video views have a stronger impact on an influencer's annual income compared to the quantity of subscribers. To express this finding in technical language for a research paper on business analytics, you can use the following statement:

"In our comprehensive analysis, it has been ascertained that video views and the quantity of subscribers exhibit notable influences on the annual income of influencers or advertisers. Notably, video views demonstrate a higher degree of significance and impact on an influencer's yearly earnings, emphasizing the pivotal role of content viewership metrics in income generation within the influencer marketing domain."

This statement communicates the key findings in a formal and technical manner, highlighting the relative importance of video views in comparison to subscribers in influencing an influencer's income.

REFERENCES

- https://www.sciencedirect.com/science/article/pii/S1386505622000466?casa_token=7hf8n_vFsfAAAAAA:5VSxCP-LF1a4hycPqUSsy2N6l34sQ_HUQwZZdf80H7wTFP18Kltos0udOV-or3xRFXMh3OBEI-nMgg
- <https://dl.acm.org/doi/abs/10.1145/2647868.2654920>
- <https://dl.acm.org/doi/abs/10.1145/2661714.2661725>
- <https://ieeexplore.ieee.org/abstract/document/7431452/>
- https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s11042-009-0309-5&casa_token=vleFlz-8KoAAAAA:PalkO2OA6Xwfowg_qhNEO7pZqReHIFbzWYvgAJJt64JALgJD7v1IYM1gtivi9yh7CGuXvNru0tXU6KqE8
- <https://www.diva-portal.org/smash/record.jsf?pid=diva2:1593439>
- [https://www.ajodo.org/article/S0889-5406\(16\)00033-0/abstract](https://www.ajodo.org/article/S0889-5406(16)00033-0/abstract)
- <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-1681.2010.05376.x>
- http://people.duke.edu/~rnau/notes_on_linear_regression_analysis--robert_nau.pdf
- https://links.sharezomics.com/assets/uploads/files/1600247928973-from_slack_logmodels2.pdf
- <https://www.jstor.org/stable/120052>
- <https://www.sciencedirect.com/science/article/pii/S0377221789904311>
- <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=63167c5dbb9bae6f0a269237a9b6a28fa7e1ac20>
- <https://dl.acm.org/doi/abs/10.1145/2959100.2959190>
- <https://www.sciencegate.app/document/10.1108/jhr-06-2020-0229>
- https://books.google.com/books?hl=en&lr=&id=xd0tNdFOOjC&oi=fnd&pg=PR7&dq=research+papers+on+linearregression&ots=dV9tztHAOS&sig=8d8HJX8qM3nKvQbzWYOM_wsZXE8
- <https://books.google.com/books?hl=en&lr=&id=745QDwAAQBAJ&oi=fnd&pg=PP2&dq=research+papers+on+linearregression+on+youtube+data&ots=5ERttQU3Mu&sig=bng0TOIGmuyquyhPRStQSiV6IBk>
- <https://www.hindawi.com/journals/complexity/2021/5599655/>
- <https://www.sciencegate.app/document/10.13189/ms.2021.090610>