Charting Success with AI and Statistics, The Path to a Million Subscribers

Martín Molina

February 29, 2024

Contents

Chapter 1

Introduction

1.1 Context

We are on the verge of entering a technological revolution driven by Artificial Intelligence (AI), promising to redefine the automation landscape.

AI is not just limited to enhancing existing processes; it is capable of creating new content and technologies from its own capabilities.

YouTube stands out as a interesting platform for AI integration, particularly due to its underlying algorithms that govern content distribution and user engagement.

1.2 Justification of the Study

Given YouTube's algorithm for content recommendation to audiences, understanding these factors and optimizing various processes through AI could revolutionize content creation and distribution, significantly reducing production time and cost.

The justification for this study lies in its potential to demonstrate the power of AI in the realm of digital content creation and distribution. It seeks to validate the capability of these technologies to automate aspects of the audiovisual sector. In my case, I found YouTube to be a prime example for verification, but the applications could be endless. From aiding in corporate training and education to creating advertising content, the possibilities are limitless.

By conducting a detailed analysis of the metrics that most significantly impact YouTube's algorithm, this study intends to unearth vital insights into strategies for content optimization.

1.3 Objectives

The primary aim of this thesis is to develop an AI-driven, nearly fully automated YouTube channel. The specific objectives are as follows:

- Create a YouTube channel where content creation is predominantly automated through AI technologies.
- Subtly modify videos in ways that influence the audience to boost various types of interactions, such as likes, comments, shares, and audience retention.
- Conduct an analysis of YouTube's metrics to identify those with the most significant influence on the platform's recommendation algorithm.
- Assess the growth of the channel and changes in audience engagement resulting from the AI-driven modifications.

This project endeavors not only to contribute to the academic field of AI in digital media but also to provide practical insights for content creators seeking to leverage AI for enhanced online presence and audience engagement.

Chapter 2

Theoretical Framework

All the information presented in the following section is the result of extensive research into YouTube's algorithm, utilizing resources provided by Google's team to all creators, combined with my own hypotheses based on the outcomes of the project.

2.1 YouTube's Algorithm: An Overview

When a video is uploaded to YouTube, it is initially shown to a small sample of viewers, and based on a series of parameters, it is decided whether to further promote the video or not. This decision is made using a complex and comprehensive algorithm, commonly known as the YouTube algorithm.

Understanding YouTube's algorithm is crucial for grasping how video recommendations work on the platform. Comprehending its operation can significantly enhance the boost that the platform can provide to a creator by adjusting a series of parameters or improvements, aiming to perfect the aspects that are most valued by the algorithm.

2.2 Performance Metrics on YouTube

Some of the most important metrics, that are crucial in understanding how the platform measures and rewards content performance, are the following:

- **Engagement Metrics:** Includes likes, comments, shares, and the overall watch time. These metrics indicate how viewers are interacting with the content.

- **Click-Through Rate (CTR):** Measures the frequency at which viewers click on a video after seeing its thumbnail, reflecting the effectiveness of the video's title and thumbnail.
- **Watch Time:** The total duration viewers spend on a video, a critical factor in assessing viewer retention and engagement.
- **Content Relevance:** The alignment of video content with the viewer's search queries and interests, analyzed through metadata like titles, descriptions, and tags.
- **Video Quality:** Considers factors like resolution, sound quality, and overall production value.
- **User Feedback:** Includes direct responses from viewers, influencing how content is perceived and ranked by the algorithm.

Each of these metrics plays a specific role in how the algorithm evaluates and prioritizes content, making them essential considerations for content creators.

2.3 YouTube Shorts: A Different Approach

YouTube Shorts introduces a slightly different approach to content creation and audience engagement. Our study's goal will be to analyze how different metrics behave when uploading shorter content on YouTube. We aim to develop a formula or strategy for content creators to optimize their presence on the platform. Special attention will be given to:

- **Viewer Engagement:** How Shorts engage viewers differently from longer content.
- **Content Coverage:** The percentage of viewers who choose to watch a Short and its impact on overall visibility.
- **Viewer Retention:** Analyzing the percentage of the video watched by viewers and its implications for content strategy.
- **Consistency:** Exploring how consistent publishing affects Shorts' performance.

This comparative analysis aims to uncover insights into optimizing content for Shorts and leveraging their unique format to enhance audience reach and engagement.

Chapter 3

System Development

3.1 Channel Creation and Conceptualization

In the initial phase of the project, I focused on creating the channel and conceptualizing its theme. Orienting the channel in terms of content is a task that requires considerable attention. Our aim is to create a channel that can be easily automated yet capable of having a future trajectory, one that can expand once a community is formed. We are looking for content that can maintain a straightforward approach, without notable variations, but has a sufficiently broad foundation so that, at a certain point, it can be expanded to include new projects and content. We must consider the implementation of AI tools, as the content should be able to be created almost automatically. At our disposal, we have *Chat GPT* for brainstorming ideas, and we can create images or voices, using them responsibly.

Finally, the decision was to name the channel *Choizy*, symbolizing the power of choice. It was a name that no one else had taken, and it allowed me to create batch content about the *Would You Rather* game. Furthermore, there's potential to later create longer videos about choices, or even develop a game, merchandise, etc. However, these expansions are not part of this project.

3.2 Integration of Artificial Intelligence in Content Creation

In the 'Integration of Artificial Intelligence in Content Creation' phase of our project, we embraced a suite of AI tools to develop the content creation process for the channel.

- *Namelix* was used for inspiration in naming the channel. This AI-powered tool is a good complement for brainstorming name ideas, providing brandable name suggestions.
- The channel's visual identity was designed using *DALL · E3*. Subsequently, AI tools provided by Canva were utilized to obtain a design without a background for vectorization purposes. The exact *prompt* provided (in Spanish) is the following:

> *Foto de un logo redondeado con una letra C en 3D. Las flechas estÃąn integradas directamente en los puntos donde la C comienza y termina su curva.*

- *Canva* also played an essential role in video production, particularly in facilitating the efficient creation of content in batches. Its user-friendly interface and extensive features allows us to produce a lot of high-quality videos at a time.
- The content of our *Would You Rather* game videos was enriched using *ChatGPT*. This AI tool turned out very interesting in the making of a lot of game options, automating all the process.
- To enhance the functionality provided by *Chat GPT*, I integrated *Power Automate*. This tool complemented *Chat GPT* by taking the raw output of game choices and organizing them into well-structured tables in Excel.
- Lastly, *ElevenLabs* played a crucial role in voice generation for our videos. This advanced AI voice synthesis technology provides a range of realistic and engaging voices, which completes the final component of the system.

Figure 3.1: Choizy's raw logo from DALL · E3.

3.3 Content Strategy and Focus

Our approach will be to publish a short video every day at various times, experimenting with those hours commonly recommended in forums as well as more unconventional times for comparison. Furthermore, we plan to make slight variations in the videos to stimulate different metrics: likes, shares, subscriptions per video, view percentage, and coverage.

The strategy is to be consistent over a 40-day period, creating and uploading an almost automated video daily, analyzing the results, and testing various metrics alongside the algorithm as a whole. This methodical approach is designed to provide insights into the optimal content strategy and the most effective ways to engage with YouTube's algorithm.

It is well known that the accuracy of empirical statistics tends to improve with the number of trials. A sample size of 40 individuals may not be large enough to guarantee definitive results. However, it will provide us with a preliminary understanding of which criteria have the most and least impact. This approach will be supported by statistical and probabilistic results, allowing us to draw informed conclusions about the effectiveness of different strategies.

Chapter 4

Results

In this chapter I want to present a comprehensive overview of the results achieved by the YouTube channel *Choizy* over approximately four months of activity. The focus here is on presenting the empirical data collected during this period, offering an initial insight into the channel's growth and performance before delving into the more detailed statistical analysis in the subsequent chapter.

4.1 Overall Channel Performance

The first video on the channel *Choizy* was uploaded on October 10, and the last one on November 28. As of February 20, approximately four months later, the channel has achieved a remarkable milestone of over **700,000 subscribers** and around **40 million views**. According to YouTube's metrics, the Revenue Per Mille (RPM) for short videos is significantly lower, standing at a fixed rate of about $0.03 per thousand views. This means that, if the channel were to be monetized, it would have generated around $1,200 during this period, in which we remember that videos were only uploaded for about 40 days. The most viewed videos on the channel have received totals of 8.3 million, 7 million, 5.8 million, and 4.6 million views each.

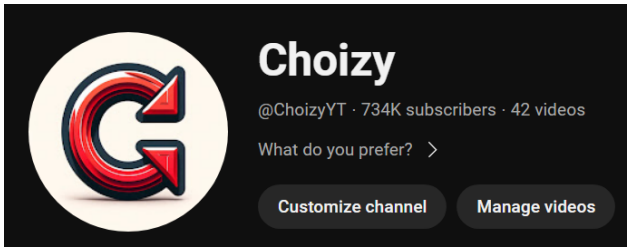


Figure 4.1: *Choizy's* channel February 20th.

In videos where I endeavored to boost interaction through likes, comments, or other means, the outcome was notably successful. This has provided an excellent basis for more in-depth statistical analysis in the subsequent section. While data such as geographical region, age, audience's content preferences, or gender might be available, I consider these metrics not particularly relevant for our analysis.

On the other hand, the analytics provide us with graphs showing the peak hours of activity of our viewers on YouTube. It could be interesting to experiment with uploading content during these peak activity hours. However, in my experience and upon reviewing various videos, the timing of uploads in the long run has not shown a significant impact on the video's performance. This finding suggests that while optimal upload times might offer some initial advantages, they do not necessarily determine a video's long-term success.

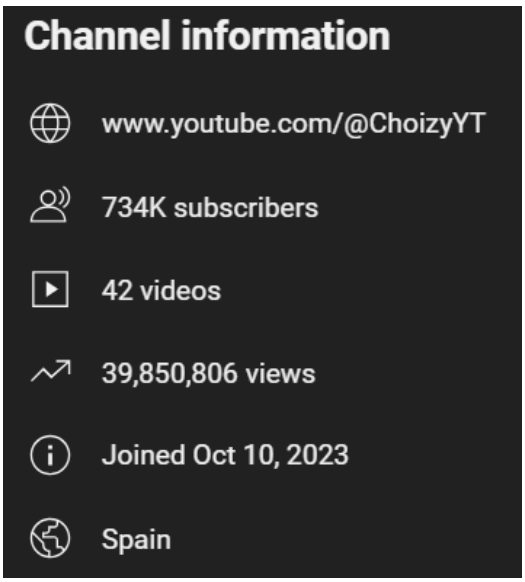


Figure 4.2: *Choizy's* empiric statistics February 20th.

4.2 Empirical Observation of Growth and Preliminary Conclusions

In the following screenshot, we can observe the performance of the channel after the first month. It was on a positive trajectory, but nothing compared to the growth it experienced a few days later, despite not uploading any new videos. Some preliminary conclusions we can draw are that consistently uploading quality content can lead to such growth. It seems that YouTube decides over time which videos to promote, and this timeframe can vary. Over time, I have observed how some

videos reached 100,000 views in just a few days, while others took months to explode in popularity, albeit to a greater extent.

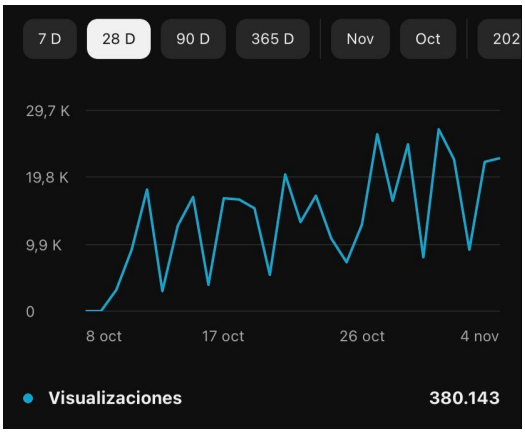


Figure 4.3: *Choizy's* growth over the first month.

Note: Detailed statistical analysis and interpretation of these results will be covered in the following chapter, providing a deeper understanding of the underlying factors driving the channel's performance.

However, it wasn't until mid-December that the channel experienced a sudden surge in growth. It gained an extraordinary number of views which significantly boosted its popularity. This surge propelled the channel from around 20,000 subscribers to approximately 200,000 in just one month. The following screenshot provides a visual representation of this fact.

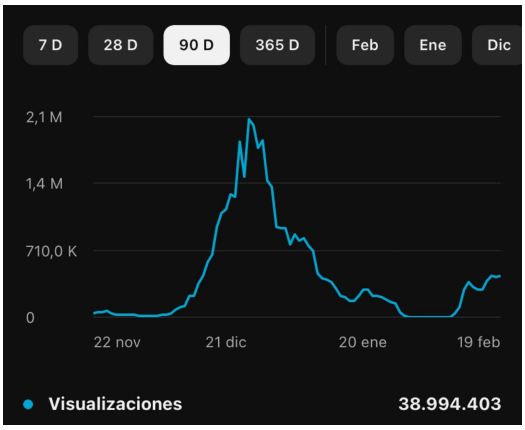


Figure 4.4: *Choizy's* sudden growth over the last months.

Chapter 5

Statistical Analysis

This chapter delves into the more technical aspects of our project, focusing on the statistical analysis of the data collected from our YouTube channel. It is important to note that this section may be more challenging to understand for those without a background in mathematical statistics or probability theory. However, I have made every effort to present the results in a reader-friendly manner, aiming to make the findings accessible to a broader audience.

For those interested in the technical details, the R scripts used for this analysis will be available in the appendix at the end of this document.

5.1 Analysis of Initial 40 Days

We divide our analysis into two phases: the initial 40 days and the global period. With this bifurcation we will try to scrutinize the statistical differences and similarities between the early stage of the channel's development and its subsequent growth.

5.1.1 Multiple Regression Analysis

One of the key methods employed in this phase of analysis is multiple regression. This approach is chosen to mitigate the issue of data inflation, which can occur when numerous variables are considered simultaneously. Multiple regression allows us to understand the relationship between several independent variables and a single dependent variable. In our case, this helps in discerning which factors most significantly impacted the channel's performance during its initial 40 days.

- **R-Squared**: The model has an R-squared value of 0.973, indicating that it explains approximately 97.3% of the variability in the visualizations. This high value suggests a good fit of the model.

- **Coefficients**:
  - **Likes**: Each additional *Like* increases the visualizations by an average of 4.3029 units. This effect is statistically significant ($p < 0.001$).
  - **Added Comments**: An increase of 5.9733 in visualizations for each additional comment, though this is not statistically significant (p = 0.348).
  - **Shared**: An increase of 37.0739 in visualizations for each time the content is shared, though this is not statistically significant ($p = 0.156$).
  - **Subscribers**: An increase of 3.7653 in visualizations for each subscriber gained. This effect is statistically significant ($p < 0.001$).
- **Interpretation**: The model indicates that "Likes" and "Subscribers" are significant predictors of visualizations, aligning with trends observed in previous models. "Added Comments" and "Shared", though present in the model, do not demonstrate clear statistical significance. Interestingly, for the *Average Viewed* and *Coverage* percentages, we did not obtain conclusive results. This could be due to the complexity of these variables or because these metrics are percentage values ranging from 0 to 100, which did not align as well with the model we used, compared to other statistics. To approach this differently, we divided the data into three groups based on percentiles and analyzed the impact of views according to these groupings.
- **Note on Multicollinearity**: The model shows a high condition number (1.95e+05), which could indicate the presence of multicollinearity. This means that some independent variables might be correlated with each other, affecting the accuracy of the estimated coefficients. A good interpretation for this is that all the interactions appear to affect in a similar way to the increase of visits, which can lead to multicollinearity. In that context, it would not represent such a big deal.

To provide an explanation for this multicollinearity, I attempted to conduct a multiple analysis by extracting different variables to observe their correlation, in addition to individual regression analyses. Removing the audience retention percentages from the analysis leads to a very significant increase in the coefficient, suggesting that **audience retention and interactions are not correlated**. However, all other metrics, such as likes, comments, shares, etc., are highly correlated, which may imply data inflation. Nonetheless, the consistency in results between individual and multiple regression models might indicate that both types of user interaction play a similar role in terms of enhancing video visibility or appeal. This hypothesis is quite logical considering the context of the analysis. It was also

found that there is no interaction between variables, which lends consistency to the notion that they might contribute similarly to a boost of the visits.

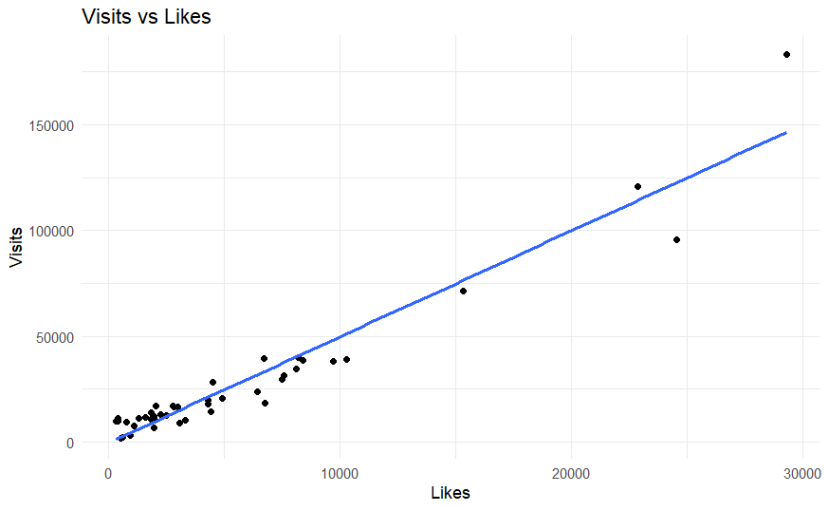


Figure 5.1: An example of Linear Regression between *Visits* and *Likes*.

5.1.2 Graphical Analysis Results for the Percentages

As we said previously, we divided the data into three groups based on percentiles. According to the graphical analysis, higher percentages generally correlate with higher average visualizations. However, as the overall percentage range is quite high, the medium-to-high category shares a lot of similarities. Therefore, in the *Average Viewed Percentage* category, there is a noticeable jump from low to medium-high. This finding indicates that a higher engagement level, as reflected by view percentages, tends to associate with increased visualizations.

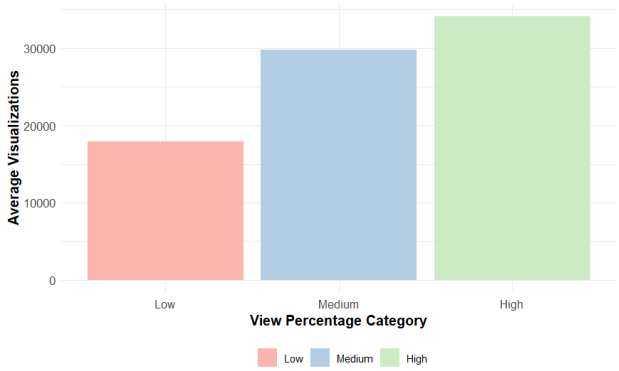


Figure 5.2: Average Visualizations by View Percentage Category.

On the other hand, *Coverage* shows a significant increase when moving from the low-medium to the high group, indicating a direct impact of this metric on the channel's performance. This trend suggests that the more viewers stay to watch the video, the higher the views tend to be.

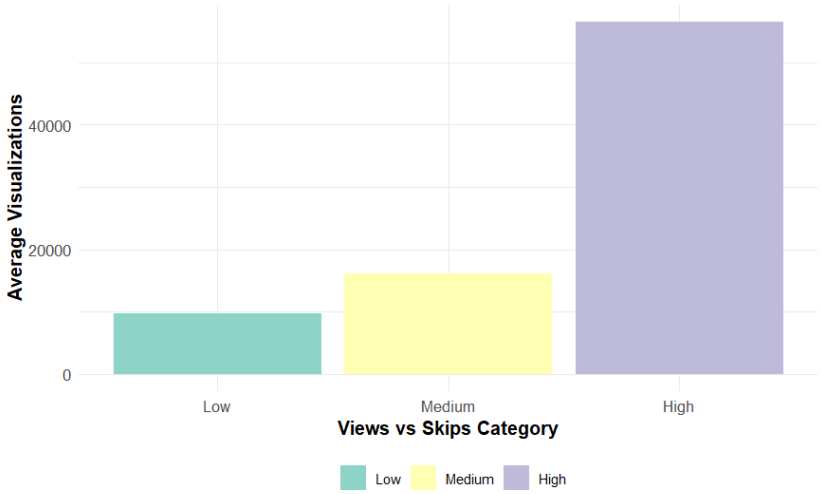


Figure 5.3: Average Visualizations by Views vs Skips Category.

Also conducted an individual correlation analysis to assess the relationship between visualizations and these two metrics.

The results of this analysis revealed a moderate to strong positive correlation of approximately 0.60 between *Views vs Skipped* and visualizations, indicating that videos with high coverage tend to have higher view counts. On the other hand, *Average Viewed Percentage* showed a weaker positive correlation of around 0.22 with visualizations, suggesting that while longer view duration have some impact on total views, other factors might play a more significant role in driving overall view counts.

For more detailed information on these analyses, a series of tables is provided in Appendix B.

5.2 Global Analysis and Comparative Review

Following the analysis of the initial 40 days, we extended our investigation to encompass the global period.

5.2.1 Multiple Regression Analysis for the Global Period

The regression analysis for the global period was conducted in a similar way as the one before.

Results:

- **Coefficients**:
  - **Likes**: Each additional *Like* increases the visualizations by an average of 2.535 units, statistically significant ($p < 0.001$).
  - **Added Comments**: This metric shows a non-significant impact on visualizations ($p = 0.42162$).
  - **Shared**: An increase of 1,766 in visualizations for each shared content, statistically significant ($p < 0.001$).
  - **Gained Subscribers**: Each additional subscriber leads to an average increase of 1.766 in visualizations, significant ($p = 0.00665$).
  - **Average Viewed Percentage**: and **Views vs Skipped** showed non-significant results.
- **Model Fit**:
  - The model's R-squared value was 0.9958, indicating it explains approximately 99.58% of the variability in visualizations.
  - The F-statistic was 1190 on 7 and 35 degrees of freedom, with a highly significant p-value ($< 2.2e - 16$).
- **Interpretation**: The model underscores the significant influence of metrics like *Likes*, *Shared*, *Gained Subscribers* on the visualizations.
- **Note on Multicollinearity**: Once again, we note a high condition number between interactions. This time, *Shares* emerge as a significant statistic. Previously, we observed an increase of 37 visits per share, which lacked significance, but now, with a larger dataset, we see a substantial rate of 1.766 visits per share, and all key metrics are around 2. This suggests that with a larger sample size, these types of metrics also influence in a similar manner, and now adding another significant statistic, which is consistent with our earlier hypothesis.

As a conclusion, upon analyzing a larger data set during the global period, we notice an intriguing pattern: the significance of *Shared* content emerges, hinting at an evolving relationship between user interactions and visualizations. This observation leads to an important reflection:

As more data is accumulated, the impact of user interactions on visualizations appears to converge around 2, with "Likes" being the most predominant feature. Furthermore, as the data set expands, more types of interactions become relevant. For instance, it's conceivable that with a greater number of videos influencing the audience to comment, we might observe a similar relationship for interactions.

5.2.2 Graphical Analysis of Percentage Metrics in the Global Study

The global study of percentage metrics yielded insightful shifts in the correlation values. Specifically, we observe a correlation of 0.47 between *Coverage* and visualizations, and a correlation of 0.4 for *Average Viewed Percentage*. This indicates a decrease in the former and an enhancement in the latter.

Most notably, the *Average Viewed Percentage* shows a significant disparity, particularly for videos with a very high percentage, indicating a substantial gap in performance. This phenomenon is clearly illustrated in the graph below, highlighting how higher percentages of viewed content correlate with fluctuations in visualizations.

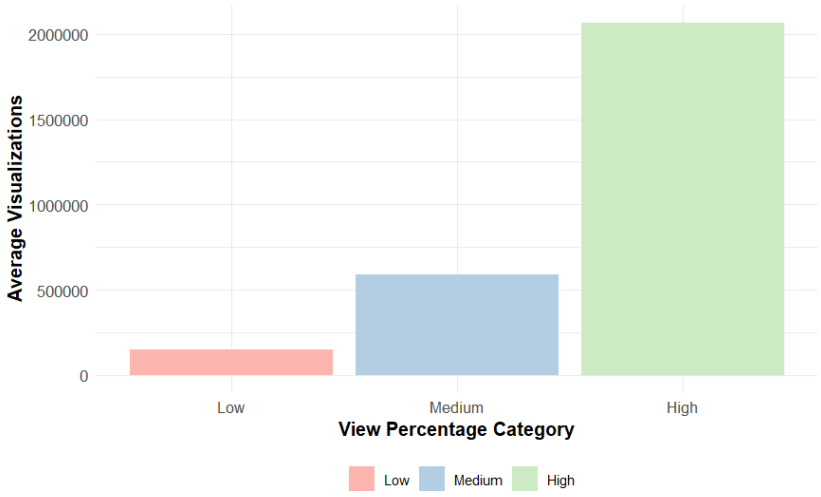


Figure 5.4: Relationship between Average Viewed Percentage and Visualizations.

For the *Coverage*'s graph, we observe a trend: videos with a low percentage of coverage tend to have significantly fewer views. This pattern underscores the importance of viewer retention in accumulating views.

Additionally, in this study, we encountered instances of videos achieving an incredible number of views. It's important to note, however, that such cases are

typically associated with higher percentages in both *Average Viewed Percentage* and *Coverage*, which suggests an implication of these factors in the Channel performance.

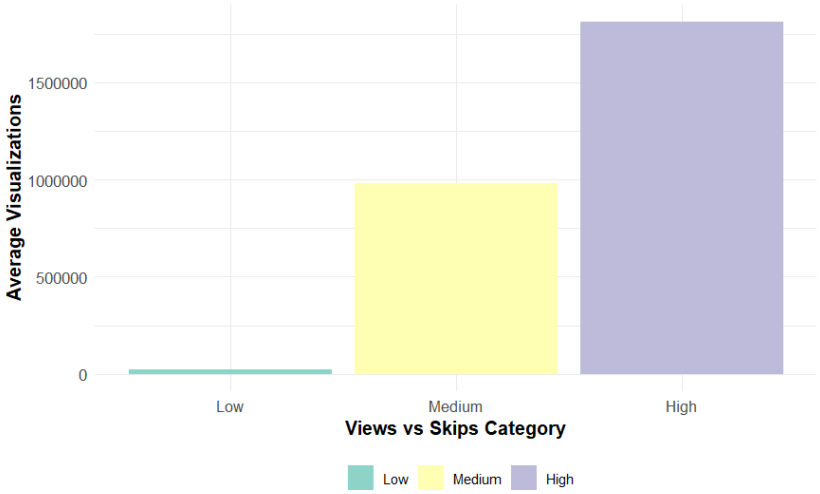


Figure 5.5: Relationship between Coverage and Visualizations.

Chapter 6

Conclusions and Recommendations

Given the results obtained in both analysis frameworks, we conclude that user interactions with videos and audience retention, both at the start and throughout the videos, are of paramount importance.

However, the differences between the two analyses lead us to conclude that initially, obtaining likes and similar interactions seemed crucial, while a high retention percentage was less so. Yet, as we gathered more data, there was a noticeable decrease in the impact of interactions on views, coupled with an significant increase of audience retention. This suggests that maintaining high retention percentages is crucial in the long term.

Observing the channel's growth, it was on a promising trajectory during the initial 40 days, but it wasn't until the second or third month that it experienced exponential growth, achieving remarkable numbers.

One interpretation of these findings is the critical importance of maintaining **consistent**, high-quality content focused on audience retention, which appears to have a more substantial influence over time. Nevertheless, efforts to boost user interactions unequivocally demonstrate a positive leap in views, indicating that they also contribute to significant growth.

From a personal perspective, the key to having popular and frequently visited content largely lies in audience retention. Predictive analysis suggests that the importance of this factor is increasing, while the impact of user interactions is diminishing. It also appears that whether it's a like, a comment, or a share, the type of interaction matters less, though likes slightly lead the way (possibly due to more available data on this metric).

Appendix A

Source Code

The following are the scripts in R added to subtract numerical data such as correlation coefficients or similar.

Note: The code snippets provided in this document are for illustrative purposes only and serve as general examples of the methods used to obtain the coefficients. They do not represent the exact code employed in my analysis.

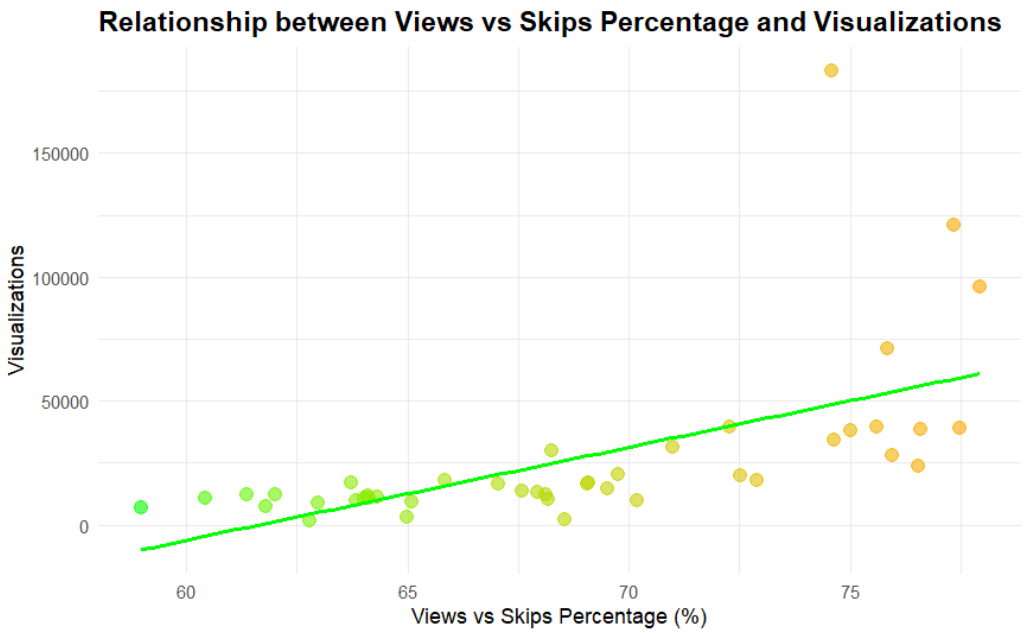
```
df <- read_excel("../Contenido␣2023-10-10_"Final Date"␣
   Choizy.xlsx")
df <- df[!is.na(df$Visualizaciones) & df$Contenido != '
   Total', ]

RM <- lm(Visualizaciones ~ `Me gusta` + `Comentarios
   aÃśadidos` + Compartido + `Suscriptores ganados` + `No
   me gusta` + `Porcentaje medio visto (%)` + `Vistos (
   frente a saltados) (%)`, data = df)
summary(RM)

cor1 <- cor(df$`Vistos (frente a saltados) (%)`, df$
   Visualizaciones, use = "complete.obs")
cor2 <- cor(df$`Porcentaje medio visto (%)`, df$
   Visualizaciones, use = "complete.obs")

cor1
cor2

df <- df[!is.na(df$Visualizaciones) & df$Contenido != '
   Total', ]
df$AvgPcg <- cut(df$`Porcentaje medio visto (%)`,
                            breaks=quantile(df$`Porcentaje
```

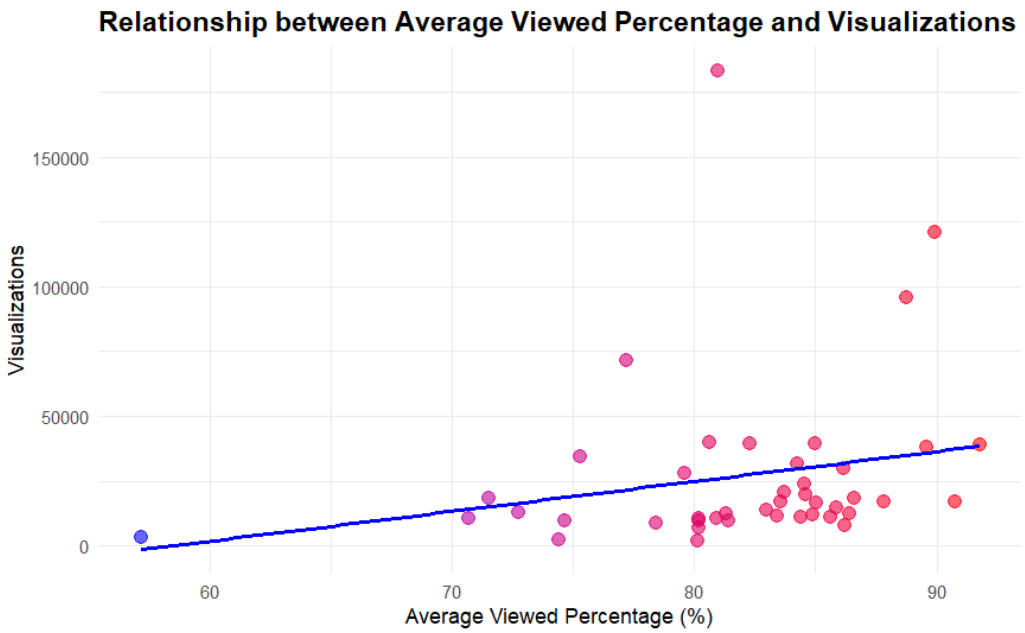
```
                                medio visto (%)`, probs=0:3/
                                3),
                              include.lowest=TRUE,
                              labels=c("Low", "Medium", "High
                                "))

df$Coverage <- cut(df$`Vistos (frente a saltados) (%)`,
                               breaks=quantile(df$`Vistos (
                                  frente a saltados) (%)`,
                                  probs=0:3/3),
                               include.lowest=TRUE,
                               labels=c("Low", "Medium", "
                                  High"))

vis-AvgPcg <- df %>%
                           group_by(Categoria_Porcentaje_
                              Visto) %>%
                           summarise(Visualizaciones_Medias =
                               mean(Visualizaciones))

vis_Coverage <- df %>%
                            group_by(Categoria_Vistos_vs_
                               Saltados) %>%
                            summarise(Visualizaciones_Medias
                               = mean(Visualizaciones))
vis-AvgPcg
vis_Coverage
```

Appendix B

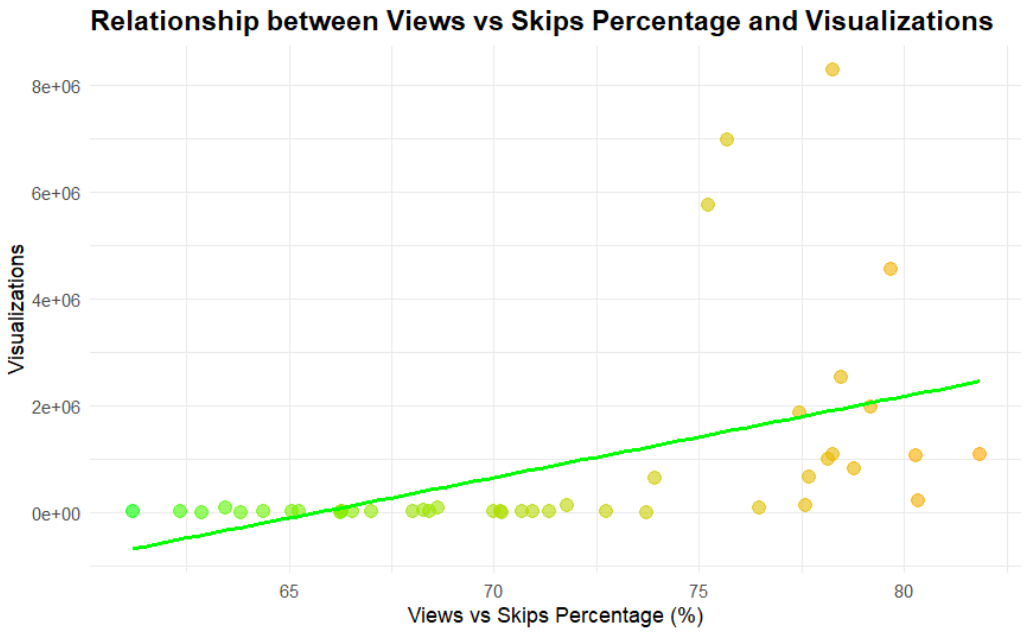
Additional Data

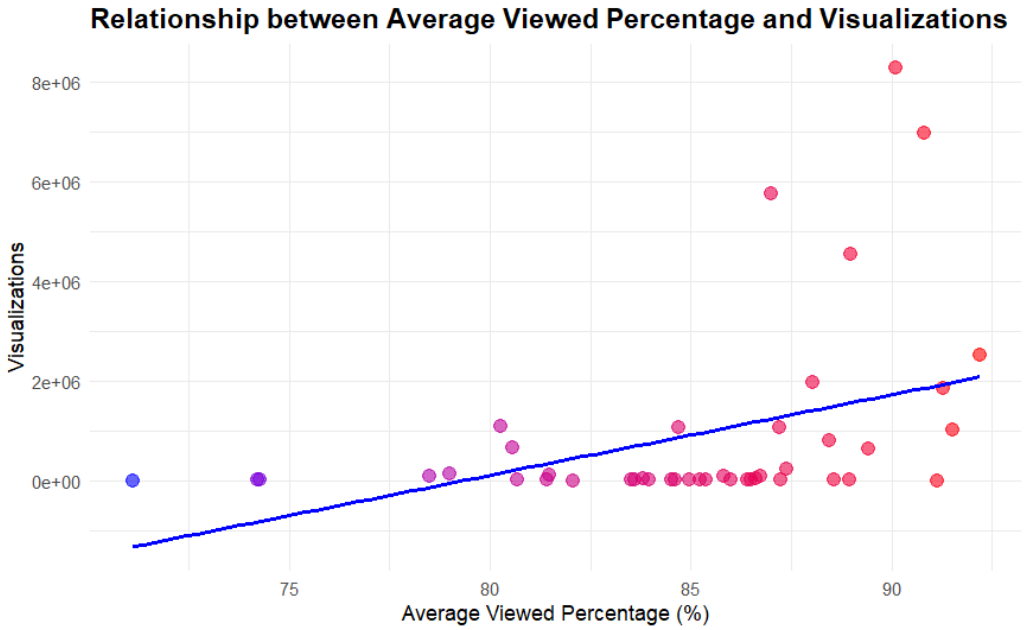
Additional graphs to first 40 days analysis:





For global analysis:




Relationship between Average Viewed Percentage and Visualizations
Visualizations
8e+06
6e+06
4e+06
2e+06
0e+00
75
80
85
90
Average Viewed Percentage (%)