King Saud University
College of Computer and Information Sciences
Information Technology Department

IT362- Data Science

CrashCourse Channel Analysis

Prepared by:

<Bashair Alsadhan, 443200668>
<Noura Alwohaibi, 443200415>
<Waref ALyousef, 442200377>
<Rama Alshebel, 443200929>
<Rana AlSayyari, 443200565>

Supervised by:
Dr. Reem Alqifari

Table of Contents

Table of Figures

Introduction

In the competitive realm of YouTube, educational channels like CrashCourse face challenges in maximizing audience reach. A key concern for CrashCourse is the inconsistency in viewership across its videos, worsened by YouTube's evolving algorithms. This project aims to analyze CrashCourse's video data, comprising 1514 observations extracted by the YouTube API, to uncover insights and strategies aimed at enhancing the channel's impact. By examining videos features such as title, views, likes, and comments, and answering five key questions, we aim to provide actionable recommendations to optimize content strategy, increase viewer engagement, and amplify CrashCourse's presence on YouTube.

In this report we will conduct the following steps from the data science process: Data Collection, Processing, Cleaning, and Exploratory Data Analysis (EDA). Throughout these steps, we will document our process, decisions, and findings. Additionally, we will conduct research on common sources of data collection bias and fairness considerations in data science. We will identify any frameworks, toolkits, or guidelines for evaluating bias and fairness in datasets. Following this, we will evaluate our collected data for potential biases, considering aspects such as representation, measurement, and historical biases. Finally, we will discuss the future steps in the data science process that we will carry out in the upcoming phases.

Bias and Fairness

1- Data bias and fairness

Despite the perceived objectivity of data analysis and machine learning, an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers, where biases creep in and undermine the accuracy of results. With the increasing prevalence of data-driven decision making, the issues of data bias and fairness have gained prominence in recent years.

Data bias refers to the systematic errors that occur in data, leading to inaccuracies or unfairness in analysis and decision-making. Bias can arise from various sources, such as the data collection, processing, or analysis methods used. Ultimately, it can result in unfair outcomes for certain groups of people and incorrect or contradictory results, therefore it is essential to detect and address data bias to ensure that data-driven analysis and decision making is fair, accurate and informatory.

2- CrashCourse's video dataset's potential biases

By focusing on analyzing a single channel and using Youtube's API as a data source, we evade some biases due to its systematic data gathering approach. However, it is not fully immune. Some biases our dataset might face include:

- **Omitted variable bias:** To protect user's privacy, Youtube's API doesn't allow some attributes such as dislikes, views over time, and favorite Count (number of users who added the video to their favorites list) to be disclosed. With omitted variable bias, the lack of a variable might affect the legitimacy of the statistic, and the effect of excluding such variables depends, even in a simple case, on a host of factors.[1] [2]
- **Information bias[3]:** it is any systematic difference from the truth that arises in the collection, recall, recording and handling of information in a study. YouTube has been known to closely monitor invalid traffic encompassing view manipulation, utilization of services to boost traffic, views from bots, and actions executed by unreal users.[4] YouTube employs both automated systems and manual data processing by specialists, thus creating possible **Algorithmic bias[5]** (systemic and repeatable errors in a computer system that create unfair outcomes) and human error.
- **Echo Chamber Bias[6]**: People tend to subscribe to channels that reinforce their existing beliefs. If you rely on data from a limited set of channels, the results might not reflect the diversity of perspectives on a topic.
- **Content Moderation and Survivorship Bias[7]**: YouTube's content moderation system might remove certain videos, even if they hold valuable information. The channel itself

[1] Clarke, K. A. (2005). *The Phantom Menace: Omitted Variable Bias in Econometric Research. Conflict Management and Peace Science*, 22(4), 341-352. https://doi.org/10.1080/07388940500339183

[2] Steiner, P. & Kim, Y. (2016). *The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases* . Journal of Causal Inference, 4(2), 20160009. https://doi.org/10.1515/jci-2016-0009

[3] Catalogue of bias collaboration. Bankhead CR, Spencer EA, Nunan D. Information bias. In: Sackett Catalogue Of Biases 2019. https://catalogofbias.org/biases/information-bias/

[4] Google. (n.d.). Definition of invalid traffic. Google AdSense Help. https://support.google.com/adsense/answer/16737?hl=en&ref_topic=9886078&sjid=16146923351955517338-EU

[5] Lee, N., Resnick, P. & Barton, G., 2019. *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*, Brookings Institution. United States of America. Retrieved from https://policycommons.net/artifacts/4141276/algorithmic-bias-detection-and-mitigation/4949849/ on 25 Mar 2024. CID: 20.500.12592/k29pdg.

[6] Brown, Megan and Bisbee, James and Lai, Angela and Bonneau, Richard and Nagler, Jonathan and Tucker, Joshua Aaron, *Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users* (May 11, 2022). Available at SSRN: https://ssrn.com/abstract=4114905 or http://dx.doi.org/10.2139/ssrn.4114905

[7] Samuels, R., Taylor, J. E., & Mohammadi, N. (2020). *Silence of the Tweets: incorporating social media activity drop-offs into crisis detection. Natural Hazards*. doi:10.1007/s11069-020-04044-2

might also delete some of its content for various reasons. This leads to CrashCourse's data being impacted by **survivorship bias**, and it occurs because we're only analyzing videos that have survived both YouTube's content moderation process and the channel's own decision to keep videos live. This can skew the analysis towards videos with specific characteristics that may not be representative of the entire content library.

- **Audio-centric Bias**[8]: The YouTube API primarily focuses on textual data. This neglects information conveyed solely through audio, potentially missing valuable insights from non-descriptive videos or those with limited text information.
- **Comment Bias**[9]: The API might return data based on comment volume or sentiment. This can favor videos that spark controversy or strong emotions, neglecting content with less active comment sections.
- **Algorithmic Drift**[10]: YouTube's recommendation and search algorithms are constantly evolving. Data collected over time might not be directly comparable due to changes in how the platform prioritizes content.
- **User Demographics Bias**[11]: The API can't directly access user demographics, but biases in who uses YouTube and how they engage with content can influence the data. For instance, data based on view counts might be skewed towards demographics with more leisure time or easier internet access.
- **Seasonality Bias**[12]: Content popularity can fluctuate based on time of year or current events. Data collected during specific seasons or holidays might not represent broader trends.

---

[8] Lai, A., Bisbee, J., Nagler, J., Tucker, J. A., & Brown, M. A. (2022, October 13). *Echo chambers, rabbit holes, and ideological bias: How YouTube recommends content to real users*. Brookings. https://www.brookings.edu/articles/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/

[9] *Comments | YouTube Data API | Google for Developers*. (n.d.). Google for Developers. https://developers.google.com/youtube/v3/docs/comments

[10] (n.d.). Deep Neural Networks for YouTube Recommendations. In *https://research.google.com/pubs/archive/45530.pdf* . Retrieved March 29, 2024, from https://static.googleusercontent.com/media/research.google.com/ar//pubs/archive/45530.pdf

[11] *Dimensions | YouTube Analytics and Reporting APIs | Google for Developers*. (n.d.). Google for Developers. https://developers.google.com/youtube/analytics/dimensions

[12] G. A. (2019, November 8). *What is seasonality and why is it important?* YouTube. https://www.youtube.com/watch?v=1kt6ZH_WJXk

3- Implications

The biases we've identified in our project using the YouTube API can significantly impact the fairness and reliability of our conclusions regarding CrashCourse's viewership and content strategy. Here's a breakdown of the implications:

1. **Inaccurate Viewership Analysis:**

   Missing Data Bias: Without access to dislikes, favorites count, and views over time, our analysis of factors like likes and comments might be misleading. We might overestimate the impact of comments or likes compared to actual viewership patterns.

2. **Unrepresentative Recommendations:**

   Algorithmic Bias & Echo Chamber Bias: The data might be skewed towards videos popular within a specific audience segment due to YouTube's recommendation system. Our recommendations to optimize content might not resonate with the broader audience CrashCourse aims to reach.

3. **Limited Understanding of Engagement:**

   Content Moderation Bias & Audio-centric Bias: Missing information due to content moderation and a focus on textual data can lead to a narrow understanding of viewer engagement. Important insights from removed videos or audio-centric content (e.g., humor, storytelling) might be overlooked.

4. **Inconsistent Viewership Patterns:**

   Comment Bias & Seasonality Bias: Data skewed towards videos with high controversy or collected during specific seasons might not reflect typical viewership patterns. Recommendations based on such data might not be effective year-round or for all types of CrashCourse content.

These biases can lead to inaccurate conclusions about what factors truly drive viewership and engagement for CrashCourse videos. The recommendations we develop based on this data and the implications might mitigate these biases.

4- Recommendations

The YouTube API is a valuable tool, but it's important to consider potential biases when using it to analyze CrashCourse's video performance. Here are some recommendations to mitigate these biases in data collection and analyzing efforts:

1- **Acknowledge limitations:** Be transparent about the limitations of our data and potential biases in our analysis. Discuss the level of confidence we have in our conclusions.
2- **Control variables**: When analyzing factors influencing viewership, account for potential confounding variables like video topic category, upload date, or video length. This isolates the specific effects of the variables we're interested in.
3- **Focus on direction of relationships:** Even with missing data, our analysis might still reveal the direction of relationships between variables. For instance, we might find that longer titles tend to correlate with slightly higher views on average.
4- **Qualitative analysis:** Consider including qualitative elements like the meaning behind the titles, this can provide valuable insights into audience preferences and engagement that might not be captured by quantitative metrics alone.
5- **Collect Data Over Time:** Longitudinal data collection allows us to track trends and identify seasonal variations in viewership. This helps differentiate content popularity from temporary spikes due to current events.
6- **Document Data Collection Methods:** Clearly document the specific methods used to collect data from the YouTube API. This allows for reproducibility and helps others understand the potential biases present in our data.

By acknowledging the biases and taking steps to implement these recommendations and mitigate them, we can increase the fairness and reliability of our conclusions, leading to more actionable recommendations for the CrashCourse channel.

Data Processing and Cleaning

Data preprocessing and cleaning represent fundamental stages in analyzing data across various domains, including scrutinizing YouTube channels like CrashCourse. Within this section, we navigate through the detailed process of refining raw data sourced from CrashCourse's channel to facilitate insightful analysis. Encompassing a diverse array of methods such as normalization, null checking, and transformation, data preprocessing assumes a pivotal role in guaranteeing the accuracy, coherence, and relevance of the data for analytical endeavors.

The Figure 1 illustrates the raw, unprocessed dataset before undergoing any data preprocessing steps:

	Title	ID	Published_date	Tags	Views	Likes	Comments	Duration	Captoins
0	Why Your Cat Looks Like That: Genetics: Crash ...	YnJPbphsoMY	2024-02-20T17:00:21Z	['vlogbrothers', 'Crash Course', 'crashcourse'...	32963	1416	24	PT11M48S	True
1	Black American History Arts & Culture | Compil...	bfH3fkIsc5U	2024-02-16T16:30:06Z	['vlogbrothers', 'Crash Course', 'crashcourse'...	16596	666	27	PT1H13M13S	True
2	Why Are All Humans Unique? Meiosis: Crash Cour...	pj1oFx42d48	2024-02-13T17:00:39Z	['vlogbrothers', 'Crash Course', 'crashcourse'...	45926	1394	36	PT12M50S	True
3	Mitosis and the Cell Cycle: Crash Course Biolo...	skPOXcVvS5c	2024-02-06T17:00:44Z	['vlogbrothers', 'Crash Course', 'crashcourse'...	47789	1338	14	PT11M11S	True
4	Photosynthesis: The Original Solar Power: Cras...	-ZRsLhaukn8	2024-01-30T17:00:00Z	['vlogbrothers', 'Crash Course', 'crashcourse'...	52459	1185	26	PT13M4S	True

Figure 1 – data frame

Figure 3 illustrates the count of columns and rows in the dataset, providing an overview of its dimensionality. Meanwhile, Figure 2 displays the data types of each column in the dataset, enabling assessment for necessary type conversions.

Get the number of rows and columns

```
num_rows, num_cols = df.shape

print("Number of rows:", num_rows)
print("Number of columns:", num_cols)
```

```
Number of rows: 1515
Number of columns: 9
```

Figure 3 - number of columns and rows

Prints the data types of each column

```
column_types = df.dtypes

print("Types of columns:")
print(column_types)
```

```
Types of columns:
Title             object
ID                object
Published_date    object
Tags              object
Views              int64
Likes              int64
Comments           int64
Duration          object
Captoins            bool
```

Figure 2- data types

1- Checking for null values

Subsequently, we execute the following code in Figure 4 to inspect the dataset for any null values, confirming that there are no missing values present within the dataset.

```
nulls_exist = df.isnull().any().any()

if nulls_exist:
    print("There are null values in the DataFrame.")
else:
    print("There are no null values in the DataFrame.")
```

```
There are no null values in the DataFrame.
```

Figure 4 - checking for null values

2- Checking for duplicated rows

Following that, in Figure 5 we proceed to check for duplicated rows by examining for any duplicated IDs within the dataset.

Checking for duplicated values in ID

```
1  if df['ID'].duplicated().any():
2      print("Duplicate values found.")
3  else:
4      print("No duplicate values.")
```

```
No duplicate values.
```

Figure 5 - checking for duplicated rows

3- Transformation

In the process of preparing the data for analysis, several transformations were implemented. Firstly, in Figure 6, Boolean captions underwent conversion into binary values, with 'True' represented as 1 and 'False' as 0. Following this, in Figure 7, the 'Published_date' column was formatted into datetime objects and subsequently converted into a specific string format. Lastly, as shown in Figure 8, the duration column was transformed into seconds. These transformations were essential in standardizing and refining the data, ensuring its compatibility with subsequent analysis procedures.

Converted boolean captions to binary values (1 for True, 0 for False)

```
df['Captoins'] = df['Captoins'].map({True: 1, False: 0})
df.head()
```

Figure 6 - converted boolean captions to binary values

Formatted the 'Published_date' column to datetime objects and then converted it to a specific string format

```
from dateutil import parser

df['Published_date'] = pd.to_datetime(df['Published_date'])
df['Published_date'] = df['Published_date'].dt.strftime("%Y-%m-%d %H:%M:%S")

df.head()
```

Figure 7 – transforming the Published-date column

Transforming duration into seconds

```
df['Duration'] = pd.to_timedelta(df['Duration']).dt.total_seconds()
df.head()
```

Figure 8 - transforming duration into second

The dataset provided here showcases the transformations that have been applied to it.

	Title	ID	Published_date	Tags	Views	Likes	Comments	Duration	Captoins
0	Why Your Cat Looks Like That: Genetics: Crash ...	YnJPbphsoMY	2024-02-20 17:00:21	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002072	0.005684	0.000810	708.0	1
1	Black American History Arts & Culture | Compil...	bfH3fkIsc5U	2024-02-16 16:30:06	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.001043	0.002673	0.000911	4393.0	1
2	Why Are All Humans Unique? Meiosis: Crash Cour...	pj1oFx42d48	2024-02-13 17:00:39	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002887	0.005596	0.001215	770.0	1
3	Mitosis and the Cell Cycle: Crash Course Biolo...	skPOXcVvS5c	2024-02-06 17:00:44	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003004	0.005371	0.000473	671.0	1
4	Photosynthesis: The Original Solar Power: Cras...	-ZRsLhaukn8	2024-01-30 17:00:00	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003297	0.004757	0.000878	784.0	1

Figure 9 - dataset after transforming

4-Normalization

Next, we proceeded with the normalization of the 'views', 'likes', and 'comments' columns. Normalization was performed to bring these numerical features to a similar scale, preventing any particular feature from dominating the analysis due to its larger magnitude. In this process, we utilized min-max normalization, ensuring that all values fell within a range of 0 to 1. This approach maintains the relative relationships between the values while standardizing their scales for more effective analysis.

In Figure 10, we present the dataset after the normalization process has been applied.

	Title	ID	Published_date	Tags	Views	Likes	Comments	Duration	Captoins
0	Why Your Cat Looks Like That: Genetics: Crash ...	YnJPbphsoMY	2024-02-20 17:00:21	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002072	0.005684	0.000810	708.0	1
1	Black American History Arts & Culture | Compil...	bfH3fkIsc5U	2024-02-16 16:30:06	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.001043	0.002673	0.000911	4393.0	1
2	Why Are All Humans Unique? Meiosis: Crash Cour...	pj1oFx42d48	2024-02-13 17:00:39	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002887	0.005596	0.001215	770.0	1
3	Mitosis and the Cell Cycle: Crash Course Biolo...	skPOXcVvS5c	2024-02-06 17:00:44	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003004	0.005371	0.000473	671.0	1
4	Photosynthesis: The Original Solar Power: Cras...	-ZRsLhaukn8	2024-01-30 17:00:00	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003297	0.004757	0.000878	784.0	1

Figure 10 - datasets after normalization

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics and trends within a dataset. In this section, we will delve into various attributes of YouTube videos from the CrashCourse channel using Python and several libraries such as Pandas, NumPy, seaborn, and matplotlib. By analyzing metrics such as views, likes, comments, Tags, and video durations, we aim to uncover patterns and insights that can inform strategic decisions and optimizations for the channel. Through statistical summaries and visualizations, we will gain valuable insights into viewer engagement and content performance.

- In Figure 11 the boxplot illustrates the distribution of video durations for CrashCourse YouTube videos, with an average duration of around 11.12 minutes, which is about 667.21 seconds. Most videos cluster closely around this average duration, indicating typical content length. However, some videos extend beyond this average, indicating variability in content length. This variability offers insights into viewer preferences and content engagement, revealing potential unique content types among the outliers.

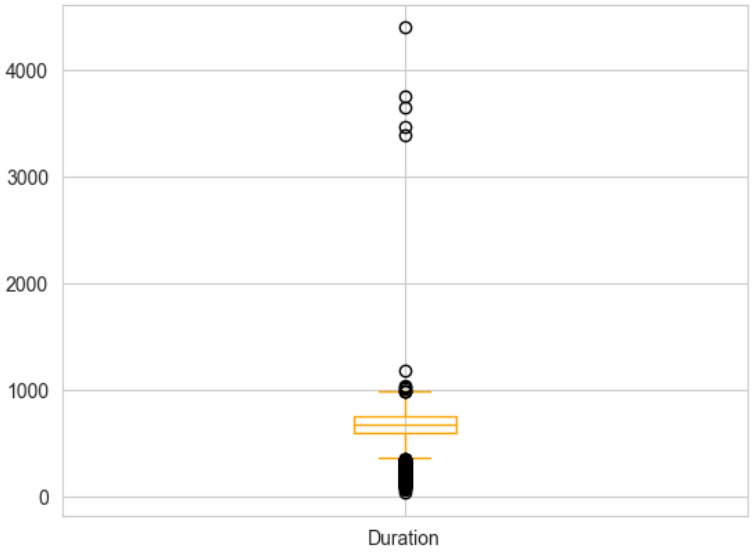


Figure 11 – boxplot of video durations

- The bar plot in Figure 12 reveals the distribution of captions for CrashCourse videos. Out of the total 1518 videos analyzed, the majority, 1481 (97.76%), have captions enabled, while only 34 (2.24%) videos do not. This indicates a strong emphasis on accessibility by CrashCourse, with a significant portion of their content being included in a wider audience.

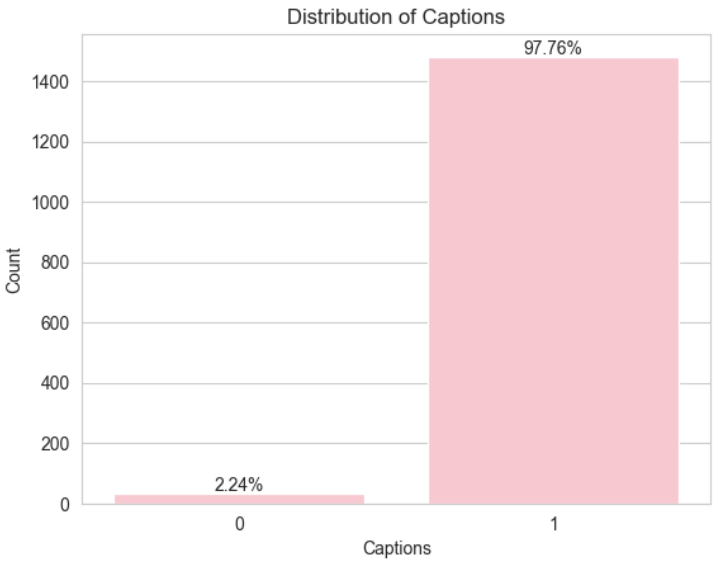
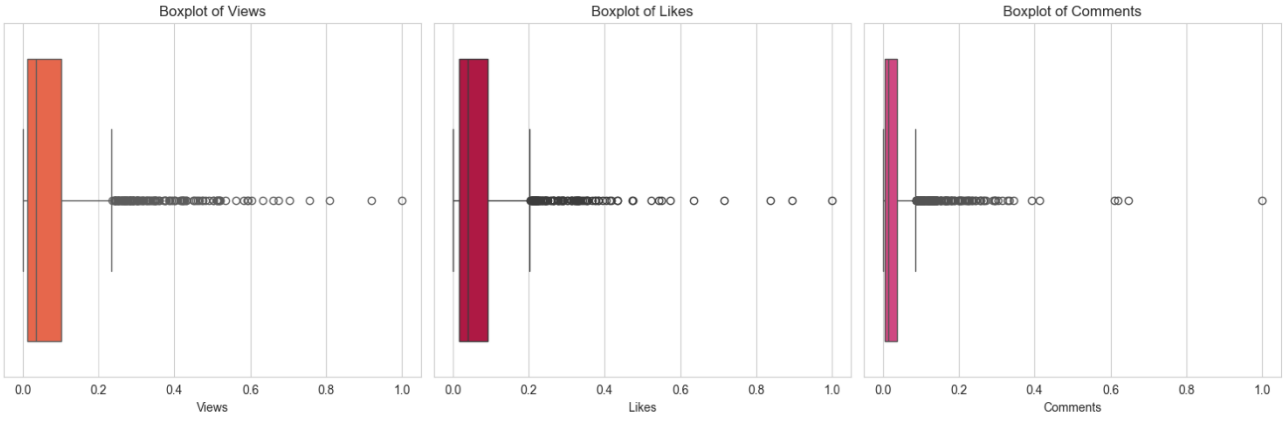


Figure 12 - distribution of captions

- In Figure 13, The subplot grid displays boxplots for views, likes, and comments. Views exhibit significant variability, with a stable median but widely extended whiskers, indicating diverse popularity levels. Similarly, likes show consistent medians but a broad range of counts, reflecting varied engagement levels. Comments also display variability, with a stable median but widely distributed counts. These plots offer insights into audience engagement patterns and content performance on the channel.

Figure 13 - boxplots for views, likes, and comments



- The correlation heatmap shown in Figure 14 visualizes the relationships between numerical attributes. It reveals a strong positive correlation between likes and views (0.94), indicating videos with more views tend to receive more likes. There's also a moderate positive correlation between comments and views (0.74), suggesting higher-viewed videos attract more comments. Interestingly, video duration shows a negligible positive correlation with the number of views (0.03), meaning that longer videos may not necessarily affect views. Captions shows weak positive correlations with views and other metrics, suggesting their minor influence on viewer engagement compared to likes and comments.

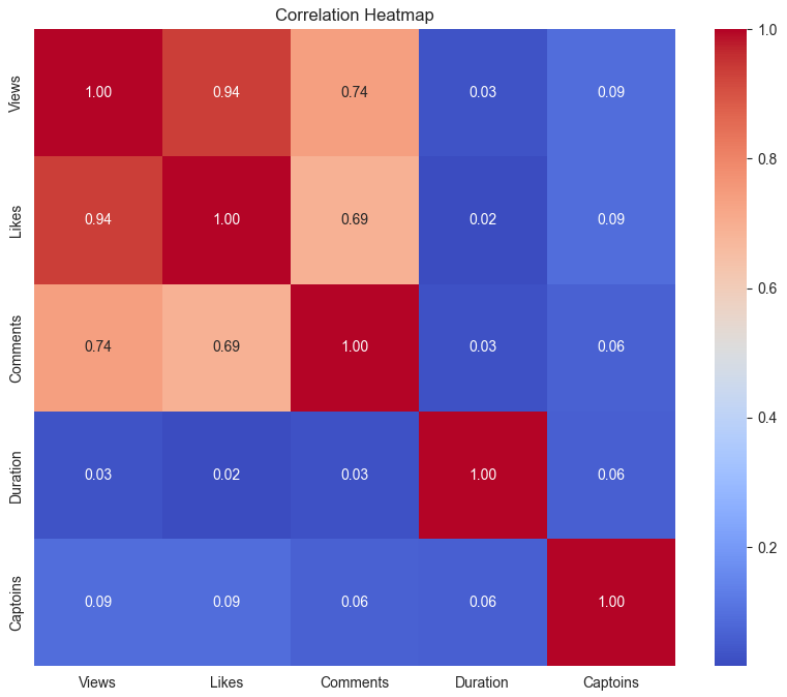


Figure 14 - correlation heatmap

- The bar shown plot in Figure 15 displays CrashCourse's top 10 most-viewed videos, revealing a strong interest in historical topics like "The Agricultural Revolution" and "Conflict in Israel and Palestine." Additionally, popular videos from psychology and biology, such as "Intro to Psychology" and "ATP & Respiration," are also featured. Overall, the diversity of subjects reflects CrashCourse's broad appeal, guiding future content decisions.

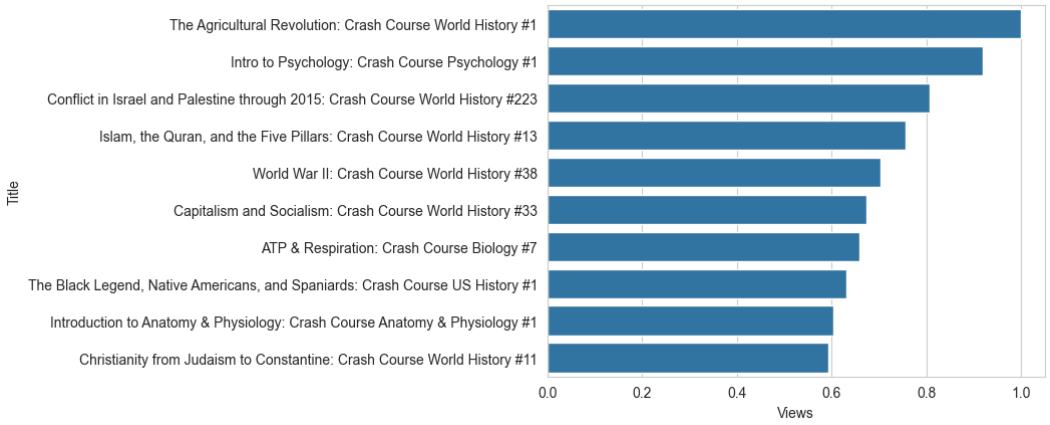
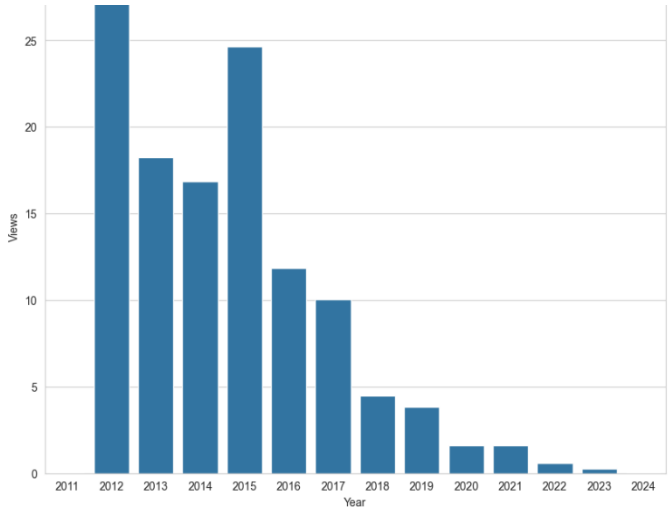


Figure 15 - top 10 most-viewed videos

- In Figure 16 bar plot displays the total views by year for CrashCourse YouTube videos. There's a declining trend in total views from 2012 to 2016, with 2012 having the highest views. Subsequently, from 2017 onwards, there's a noticeable decrease in views. This decline may result from changing viewer preferences or increased competition.

Figure 16 - total views by year



- Total views by month is shown in Figure 17 using a bar plot. Views vary throughout the year, with peaks in February, March, and April, possibly reflecting heightened interest in educational content during these months. December records the lowest views, likely due to holiday seasons. Understanding these monthly viewership trends can inform content scheduling strategies for maximizing viewer attention.

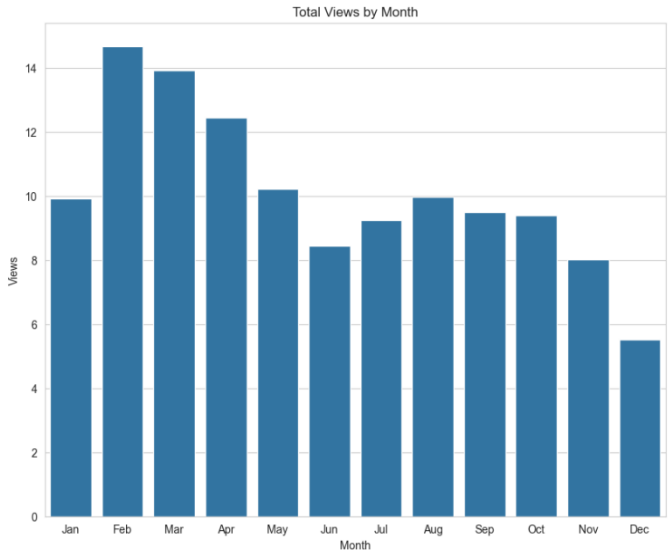
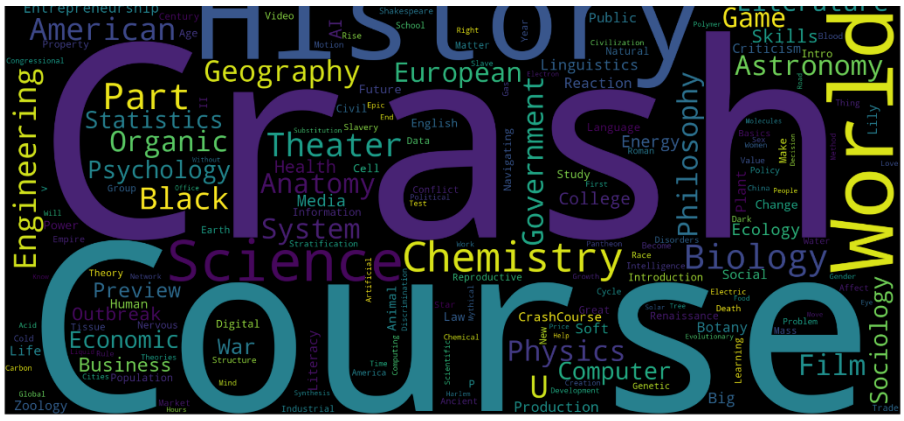


Figure 17 - Total views by month

- The word cloud visualization shown in Figure 18 captures the most frequent word occurrences in the titles. Obviously, the word "Crash" appears prominently, likely due to its association with the channel's name. Other common words may represent recurring themes or subjects covered in the videos, providing insights into the channel's content focus.

Figure 18 - word cloud visualization



- The Histogram visualization shown in Figure 19 captures the frequency of the number of tags per video. We can observe that having 6 tags is the most frequent coming over 120, While most videos have around 12 to 25 tags.

  The distribution shown in the histogram is skewed to the right, this means that most videos have a relatively few tags, with a smaller number of videos having many tags. Having right skewed means that the average (18.5) isn't representative of most videos, since it might be higher than the number of tags that most videos have. The median (17) would be a better indicator of how many tags most videos have in this case.

  This skew could stem from the tendency of people to assign just a few basic tags to most videos, but then put in more effort for specific or unusual videos. This hypothesis would explain the extremely high frequency for videos that have only 6 tags.

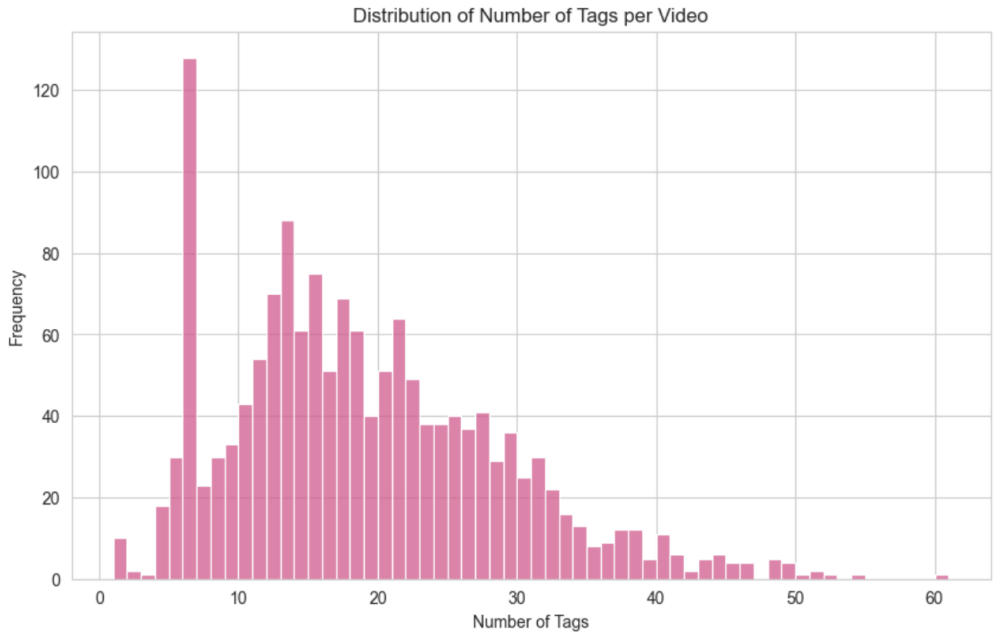


Figure 19 - Distribution of number of tags per video.

Future Steps

Following the completion of exploratory data analysis (EDA), various potential next steps emerge to further enhance the project's insights and outcomes:

1- Hypothesis Testing: Build on the patterns identified during EDA by formulating hypotheses for further investigation. For example, hypotheses could explore the relationship between video duration and viewer engagement metrics such as likes and comments.

2- Advanced Modeling Techniques: Explore the application of advanced machine learning models to predict viewer engagement or video performance based on various features. Techniques like regression analysis or time series forecasting could be employed to uncover hidden patterns and trends within the data.

3- Content Optimization Strategies: Use insights gained from our EDA to develop content optimization strategies aimed at improving viewer engagement and channel performance. This could involve experimenting with different video formats, lengths, or topics to better resonate with the target audience.

4- A/B Testing: Implement A/B testing methodologies to experiment with different content variations and measure their impact on viewer engagement metrics. This iterative approach can help validate hypotheses and refine content strategies based on empirical evidence.

References

1- Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. Conflict Management and Peace Science, 22(4), 341-352. https://doi.org/10.1080/07388940500339183

2- Steiner, P. & Kim, Y. (2016). The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases . Journal of Causal Inference, 4(2), 20160009. https://doi.org/10.1515/jci-2016-0009

3- Catalogue of bias collaboration. Bankhead CR, Spencer EA, Nunan D. Information bias. In: Sackett Catalogue Of Biases 2019. https://catalogofbias.org/biases/information-bias/

4- Google. (n.d.). Definition of invalid traffic. Google AdSense Help. https://support.google.com/adsense/answer/16737?hl=en&ref_topic=9886078&sjid=16146923351955517338-EU

5- Lee, N., Resnick, P. & Barton, G., 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, Brookings Institution. United States of America. Retrieved from https://policycommons.net/artifacts/4141276/algorithmic-bias-detection-and-mitigation/4949849/on 25 Mar 2024. CID: 20.500.12592/k29pdg.

6- Brown, Megan and Bisbee, James and Lai, Angela and Bonneau, Richard and Nagler, Jonathan and Tucker, Joshua Aaron, Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users (May 11, 2022). Available at SSRN: https://ssrn.com/abstract=4114905or http://dx.doi.org/10.2139/ssrn.4114905

7- Samuels, R., Taylor, J. E., & Mohammadi, N. (2020). *Silence of the Tweets: incorporating social media activity drop-offs into crisis detection. Natural Hazards*. doi:10.1007/s11069-020-04044-2

8- Lai, A., Bisbee, J., Nagler, J., Tucker, J. A., & Brown, M. A. (2022, October 13). *Echo chambers, rabbit holes, and ideological bias: How YouTube recommends content to real users*. Brookings. https://www.brookings.edu/articles/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/

9- *Comments | YouTube Data API | Google for Developers*. (n.d.). Google for Developers. https://developers.google.com/youtube/v3/docs/comments

10- (n.d.). Deep Neural Networks for YouTube Recommendations. In *https://research.google.com/pubs/archive/45530.pdf* . Retrieved March 29, 2024, from https://static.googleusercontent.com/media/research.google.com/ar//pubs/archive/45530.pdf

11- *Dimensions | YouTube Analytics and Reporting APIs | Google for Developers*. (n.d.). Google for Developers. https://developers.google.com/youtube/analytics/dimensions

12- G. A. (2019, November 8). *What is seasonality and why is it important?* YouTube. https://www.youtube.com/watch?v=1kt6ZH_WJXk