Abstract:

In this paper, we aim to provide a comprehensive analysis of a YouTube channel's performance, audience demographics, and video engagement levels. Using data from the channel and third-party tools, we will evaluate key metrics such as the number of subscribers, views, and comments, as well as factors that contribute to the success of videos, such as the type of content, length of the videos, and frequency of uploads. Furthermore, we will analyze the feedback received from viewers through comments and ratings to gain deeper insights into what the audience likes and dislikes. Our goal is to provide actionable recommendations that will improve the channel's performance and help it reach its full potential. The results of this analysis will be valuable for YouTube content creators and businesses who want to understand their channel's strengths and weaknesses and identify opportunities for growth and improvement.

In addition to evaluating the performance of the YouTube channel, this paper will also explore the competitive landscape and identify key trends and best practices in the niche. This will involve comparing the channel's performance with that of other channels in the same niche and analyzing the strategies used by successful channels to attract and retain subscribers. The data collected from this analysis will be used to develop a content strategy that is tailored to the specific needs and preferences of the channel's audience. This will involve identifying the type of content that resonates best with the audience, the optimal length and frequency of videos, and the most effective marketing strategies for attracting new subscribers and increasing engagement levels. The results of this paper will provide a roadmap for growth and improvement for the YouTube channel and will help it reach its full potential.

Keywords: Youtube, Channel analysis, Machine Learning

Introduction:

Founded in 2005, Youtube has grown to become the second largest search engine in the world (behind Google) that processes more than 3 billion searches per month. It is, however, generally a myth how the Youtube algorithm works, what makes a video get views and be recommended over another. In fact, YouTube has one of the largest scale and most sophisticated industrial recommendation systems in existence. For new content creators, it is a challenge to understand why a video gets video and others do not. There are many "myths" around the success of a Youtube video, for example if the video has more likes or comments, or if the video is of a certain duration. It is also worth experimenting and looking for "trends" in the topics that Youtube channels are covering in a certain niche.
The scope of this small paper is limited to data science channels and I will not consider other niches (that might have a different characteristics and audience base). Therefore, in this paper will explore the statistics of around 10 most successful data science Youtube channel.

In this paper, we will analyze the data of a YouTube channel to understand its performance and determine key metrics such as the channel's overall growth, its audience demographics, and the engagement levels of its videos. We will also compare the channel's performance with other channels in the same niche to identify strengths and areas for improvement. The goal of this analysis is to provide insights that can inform the channel's content strategy and help it grow its audience and engagement over time.

Additionally, we will be exploring the factors that contribute to the success of videos on the channel, such as the type of content, the length of the videos, and the frequency of uploads. We will also be evaluating the effectiveness of the channel's marketing strategies and the use of keywords and tags.

Furthermore, we will be looking at the feedback received from viewers through comments and ratings to gain further insights into what the audience likes and dislikes. By conducting a comprehensive analysis of the YouTube channel, we hope to provide actionable recommendations that will improve the channel's overall performance and help it reach its full potential.

Existing System:

The existing system for YouTube channel analysis typically includes manual tracking of metrics such as the number of subscribers, views, likes, and comments, as well as data obtained through third-party tools such as Google Analytics and Social Blade. These tools provide valuable insights into the channel's performance and audience demographics, but they can be time-consuming to use and may not offer a complete picture of the channel's strengths and weaknesses.

Additionally, there may be limitations on the data that can be obtained from these tools, particularly with regard to the content of the videos and the feedback received from viewers. To overcome these limitations, some YouTube content creators and businesses use custom data analytics solutions, such as those provided by data science teams or specialized marketing agencies. These solutions are designed to provide more comprehensive and in-depth insights into the channel's performance, but they can be expensive and may require specialized technical expertise to implement and use.

Proposed System:

Within this paper, I would like to explore the following:

- Getting to know Youtube API and how to obtain video data.
- Analyzing video data and verify different common "myths" about what makes a video do well on Youtube, for example:
  - Does the number of likes and comments matter for a video to get more views?
  - Does the video duration matter for views and interaction (likes/ comments)?
  - Does title length matter for views?
  - How many tags do good performing videos have? What are the common tags among these videos?
Across all the creators I take into consideration, how often do they upload new videos? On which days in the week?

• Explore the trending topics using NLP techniques

• Which popular topics are being covered in the videos (e.g. using wordcloud for video titles)?

• Which questions are being asked in the comment sections in the videos

Methodology:

1. Obtain video meta data via Youtube API for the top 10-15 channels in the data science niche (this includes several small steps: create a developer key, request data and transform the responses into a usable data format)

2. Preprocess data and engineer additional features for analysis

3. Exploratory data analysis

4. Conclusions

Datasets:

As this paper is particularly focused on data science channels, I found that not many readily available datasets online are suitable for this purpose. I created my own dataset using the Google Youtube Data API version 3.0. The dataset is a real-world dataset and suitable for the research. According to Youtube API's guide, the usage of Youtube API is free of charge given that your application send requests within a quota limit.

the selection of the top 10 Youtube channels to include in the research is purely based on my knowledge of the channels in data science field and might not be accurate.

Implementation:

SOFTWARE AND HARDWARE REQUIREMENTS:

Paper was done on Jupyter Notebook hence, no software or hardware is required. It works on any Operating System and Browser.

LANGUAGES:

• Python APPLICATIONS:
  • Anaconda
  • Spyder IDE

Excel LIBRARIES:

• Pandas
• Numpy
• YOUTUBE api
• NLTK
I noticed the count columns in channel_data is currently in string format, so I will convert them into numeric so that we can visualize and do numeric operations on them.

```python
# Convert count columns to numeric columns
numeric_cols = ['subscribers', 'views', 'totalVideos', 'playkiId']
channel_data[numeric_cols] = channel_data[numeric_cols].apply(pd.to_numeric, errors='coerce')
```

Let's take a look at the number of subscribers per channel to have a view of how popular the channels are when compared with one another.

```python
ax = sns.barplot(x='channelName', y='subscribers', data=channel_data.sort_values('subscribers', ascending=False)
ax.set_major_formatter(ticker FuncFormatter(lamba x, pos: '{:.0f}'.format(x/1000) + 'K'))
plot = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90)
```

Fig: 1 Channel Data Statistics

```python
# Write video data to CSV file for future references
videos_df.to_csv('videos_data_top10_channels.csv')
```

```
video_id | comments
---------|------------------
...      | ...              
...      | ...              
3728     | [TVMo0eA5d]      | Informational, love your content, man! Keep it up!...
3739     | [9R5N0k8q]      | Hi! I have a degree in biochemistry, but c...
3740     | [PovyLc56W]     | You seem to be encouraging me. Win! Can you guy...
3741     | [Y57U70qA]      | Can you provide us with some documentation of...
3742     | [FRBV5H]        | You told me to watch your first video, it was...
```

29
### Fig 2: Comment Dataframe

<table>
<thead>
<tr>
<th>video_id</th>
<th>channel_title</th>
<th>title</th>
<th>description</th>
<th>tags</th>
<th>published_at</th>
<th>views</th>
<th>likes</th>
<th>share</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Data Science</td>
<td>What is Alexa Rank: How to Improve Alexa Site...</td>
<td>In website traffic metrics, Alexa Rank is a rank, website traffic visitors engage...</td>
<td>Alexa rank, website traffic visitors engage...</td>
<td>2022-01-01</td>
<td>149</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Data Science</td>
<td>Sessions in Google Analytics (User Activity)...</td>
<td>In Google Analytics, sessions is a period during which a unique visitor...</td>
<td>Sessions, unique visitors, google analytics, us...</td>
<td>2021-11-17</td>
<td>802</td>
<td>8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data Science</td>
<td>Unique Visitors</td>
<td>Sessions</td>
<td>Website Traffic...</td>
<td>In Google Analytics, unique visitors are tracked, sessions, unique visitors, google analytics...</td>
<td>2021-12-17</td>
<td>333</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Data Science</td>
<td>Data Science</td>
<td>Google Analytics Events</td>
<td>Event Tracking</td>
<td>Use...</td>
<td>In Google Analytics, Events are measured as a...</td>
<td>2021-12-17</td>
<td>300</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Data Science</td>
<td>Testing and Online Experimentation</td>
<td>Join Data Science</td>
<td>and Staging for a common...</td>
<td>Test a/b testing, experimentation, Em...</td>
<td>2021-10-30</td>
<td>210</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Data Science</td>
<td>SentDex</td>
<td>How to Sort Python Dictionary by Value of Keys...</td>
<td>Python dictionaries, python dictionary sorting, sort...</td>
<td>2013-04-01</td>
<td>6585</td>
<td>497</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Fig 3: Video Dataframe respectively
Fig : 4 Comment Count & like count vs View Count

Does the number of likes and comments matter for a video to get more views?
After correcting for the absolute number of views, it turns out that the correlation is much less clear. The comment-view relationship seems to completely disappear: a lot of videos have millions of views and very few comments, while some videos have very few views have better interaction. However, it is understandable that comments take more effort than views and likes, and normally comments would die off when the video gets older.

As for like-view relationship, we can still see some positive correlation between views and like ratio (though very subtle), which means that the more views a video has, the more people would hit the like button! This seems to support the idea of social proof, which means that people tend to like better the products that are already liked by many other people.

Does the video duration matter for views and interaction (likes/ comments)?
In [ ]:

```python
sns.histplot(data=video_df[video_df['duratioSecs'] < 10000], x='duratioSecs', bins=30)
```

Out[ ]: <AxesSubplot: x label='duratioSecs', y label='Count'>

Fig: 6 duration in Secs Vs Count
Does title length matter for views?

There is no clear relationship between title length and views as seen in the scatterplot below, but most-viewed videos tend to have average title length of 30-70 characters.

```
In [ ]:
sns.scatterplot(data = video_df, x = "titlelength", y = "viewCount")
```

Out[ ]: <AxesSubplot:xlabel='titlelength', ylabel='viewCount'>

Fig: 7 Titlelength Vs Viewcount
As I'm interested to see what the creators are making videos about and which terms most frequently appear in their video titles, I will create a wordcloud for the most common words. We first need to remove the stopwords such as "you", "I", "the", etc. which do not contribute a lot to the meaning of the title.

![Wordcloud for words in title]

It can be seen that most common words are Data, Python, Tutorial, Science, Projects, Analysis, Programming, Learning, which is very expected.

Fig: 8 Wordcloud for words in title
It can be seen that most common words are Data, Python, Tutorial, Science, Papers, Analysis, Programming, Learning, which is very expected.
Views distribution per channel

With the video statistics for all channel, now we can see how the views are distributed per channel. Some channels might have a lot of views on one of their videos and the rest do not receive many views. Other channels might have more evenly distribution views per video. It can be observed that Corey Schafer, sentdex and Luke Barousse have quite large variance in their views, suggesting that they have a few viral videos. Alex The Analyst, Krish Naik and Data Science Dojo have less views overall but the views are more consistent across videos.
Which day in the week are most videos uploaded?
It's interesting to see that more videos are uploaded on Mondays and Fridays. Fewer videos are uploaded during the weekend. This could be because of the nature of the niche that is more geared towards tutorials and heavy materials, which is not suitable for weekends' consumption. But it could also just means that most creators work on their videos during the weekend or during the week and upload them beginning of the week or Friday.

**Conclusion:**
In this paper, we have explored the video data of the 9 most popular Data science/ Data analyst channels and revealed many interesting findings for anyone who are starting out with a Youtube channel in data science or another topic:
- The more likes and comments a video has, the more views the video gets (it is not guaranteed that this is a causal relationship, it is simply a correlation and can work both way).
Likes seem to be a better indicator for interaction than comments and the number of likes seem to follow the "social proof", which means the more views the video has, the more people will like it.

- Most videos have between 5 and 30 tags.
- Most-viewed videos tend to have average title length of 30-70 characters. Too short or too long titles seem to harm viewership.
- Videos are usually uploaded on Mondays and Fridays. Weekends and Sunday in particular is not a popular time for posting new videos.
- Comments on videos are generally positive, we noticed a lot "please" words, suggesting potential market gaps in content that could be filled essential.

References:


[2] Converting video durations to time
