Journal of Advanced Research in Technology and Management Sciences
Volume: 05 Issue: 04 ISSN: 2582-3078 April 2023
Available online at: http://www.jartms.org

YOUTUBE CHANNEL ANALYSIS

Dr.IMTIYAZ KHAN
Assistant Professor CS and AI dept MJCET OU Hyderabad TS India
MIR ABED ALI OSMANI
BE CSE AI DS CS and AI dept MJCET OU Hyderabad TS India
MOHD.ZAIN JAMAL
BE CSE AI DS CS and AI dept MJCET OU Hyderabad TS India

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. Prepocess 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 manyreadily 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 thatyour application send requests within a quota limit.
the selection of the top 10 Youtube channels to include in the research is purely basedon 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 hardwareis required. It works on any Operating System and Browser.

LANGUAGES :
- Python APPLICATIONS :
- Anaconda
- Spyder IDE

Excel LIBRARIESS:
- Pandas
- Numpy
- YOUTUBE api
- NLTK

●Seaborn,Matplotlib

```
channel_data
```

	channelName	subscribers	views	totalVideos	playlistId
0	Alex The Analyst	158000	5900052	126	UU7cs8q-gJRlGwj4A8OmCmXg
1	Luke Barousse	120000	5732718	68	UULLw7jmFsvfIVaUFsLs8mlQ
2	Data Science Dojo	79600	4504751	279	UUzL_0nIe8B4-7ShhVPfJkgw
3	StatQuest with Josh Starmer	645000	32254787	211	UUtYLUTtgS3k1Fg4y5tAhLbw
4	sentdex	1100000	99602241	1237	UUfzICWGWYyIQ0aLC5w48gBQ
5	Krish Naik	504000	41660941	1289	UUNU_lfiiWBdtULKOw6X0Dig
6	Tina Huang	238000	8229073	81	UU2UXDak6o7rBm23k3Vv5dww
7	Corey Schafer	879000	67492239	230	UUCezIgC97PvUuR4_gbFUs5g
8	Ken Jee	182000	5532488	221	UUiT9RITQ9PW6BhXK0y2jaeg

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.

```
# Convert count columns to numeric columns
numeric_cols = ['subscribers', 'views', 'totalVideos']
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.

```
sns.set(rc={'figure.figsize':(10,8)})
ax = sns.barplot(x='channelName', y='subscribers', data=channel_data.sort_values('subscribers', ascending=False))
ax.yaxis.set_major_formatter(ticker.FuncFormatter(lambda x, pos: '{:,.0f}'.format(x/1000) + 'K'))
plot = ax.set_xticklabels(ax.get_xticklabels(),rotation = 90)
```

Fig: 1 Channel Data Statistics

In [29]:
```
comments_df
```

Out[29]:

	video_id	comments
0	_eZRkmRfVTM	[Great course as always! Easy to follow and pi...
1	s3JmRxs53W4	[Wow Alex😍 \nYou are a saviour. Thank you!!!❤️...
2	yDG5KiIOZcQ	[Thanks!, Hi Alex, just found your channel and...
3	Z7hPEwCzk2s	[XLOOKUP is not working in my excel. How can I...
4	XRPyj7cKVsQ	[Very good video! A compact piece of knowledge...
...	...	...
3738	IrjTWNV0eAY	[Informational. Love your content, man. Keep i...
3739	RRSRKf9eQxc	[Hi Ken! I have a degree in biochemistry but c...
3740	IFceyuL6GZY	[You seem to be encouraging me. \n\nCan you gu...
3741	Y_SMU701qIA	[Can you provide us with some documentation of...
3742	qfRhKHV8-t4	[You told me to watch your first video. It was...

3743 rows × 2 columns

In [30]:
```
# Write video data to CSV file for future references
video_df.to_csv('video_data_top10_channels.csv')
comments_df.to_csv('comments_data_top10_channels.csv')
```

Fig:2 Comment Dataframe

```
In [13]: video_df
```

Out[13]:

	video_id	channelTitle	title	description	tags	publishedAt	viewCount	likeCount	favoriteCount	commen
0	iJoXTdV6C5I	Data Science Dojo	What is Alexa Rank | How to Improve Alexa Site...	In website traffic metrics, Alexa Ranks is a r...	[Alexa ranks, website traffic, visitors engage...	2022-01-06T18:33:55Z	148	3	0	
1	MUQnmOnhKMY	Data Science Dojo	Sessions in Google Analytics | User Activity |...	In Google Analytics, sessions is a period devo...	[Sessions, sessions live, google analytics, us...	2021-12-24T18:46:43Z	302	8	0	
2	KjC0IRzD3Bo	Data Science Dojo	Unique Visitors | Sessions | Website Traffic |...	In Google Analytics, unique visitors are label...	[sessions, unique visitors, google analytics, ...	2021-12-17T19:32:35Z	333	10	0	
3	5p9I0EW0X_8	Data Science Dojo	Google Analytics Events | Event Tracking | Use...	In Google Analytics, Events are measured as us...	[events, event hit, interactions, unique event...	2021-12-10T20:29:08Z	388	18	0	
4	nbvjkP8U9yg	Data Science Dojo	Testing and Online Experimentation	Join Data Science Dojo and Statsig for a conve...	[ab testing, a/b testing, experimentation, Emm...	2021-12-09T08:39:15Z	210	7	0	
...	...	...	...	...	...	...	...	...	...	
3717	MGD_b2w_GU4	sentdex	How to Sort a Python Dictionary By Value or Key!	Sentdex.com\nFacebook.com/sentdex\nTwitter.com...	[python dictionary, python dictionary sort, so...	2013-06-10T14:57:25Z	65033	459	0	
			Python's		[logging with python, python	2013-06-				

Fig 3: Video Data frame respectively

```
]: fig, ax =plt.subplots(1,2)
   sns.scatterplot(data = video_df, x = "commentCount", y = "viewCount", ax=ax[0])
   sns.scatterplot(data = video_df, x = "likeCount", y = "viewCount", ax=ax[1])
```

```
]: <AxesSubplot:xlabel='likeCount', ylabel='viewCount'>
```

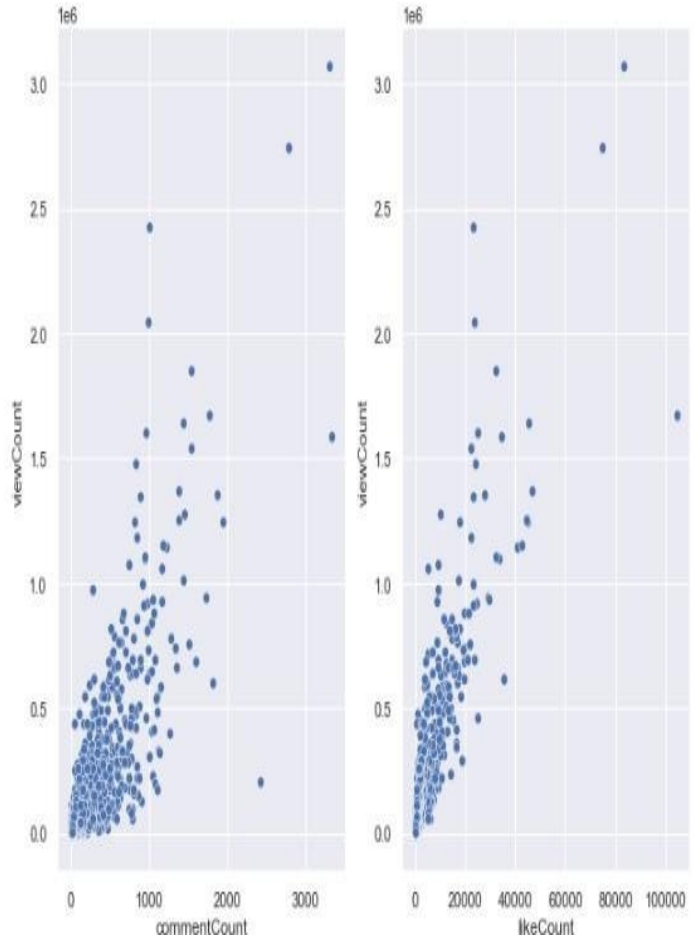


Fig : 4 Comment Count & like count vs View Count

Does the number of likes and comments matter for a video toget more views?

```
In [ ]: fig, ax =plt.subplots(1,2)
        sns.scatterplot(data = video_df, x = "commentRatio", y = "viewCount", ax=ax[0])
        sns.scatterplot(data = video_df, x = "likeRatio", y = "viewCount", ax=ax[1])
```

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

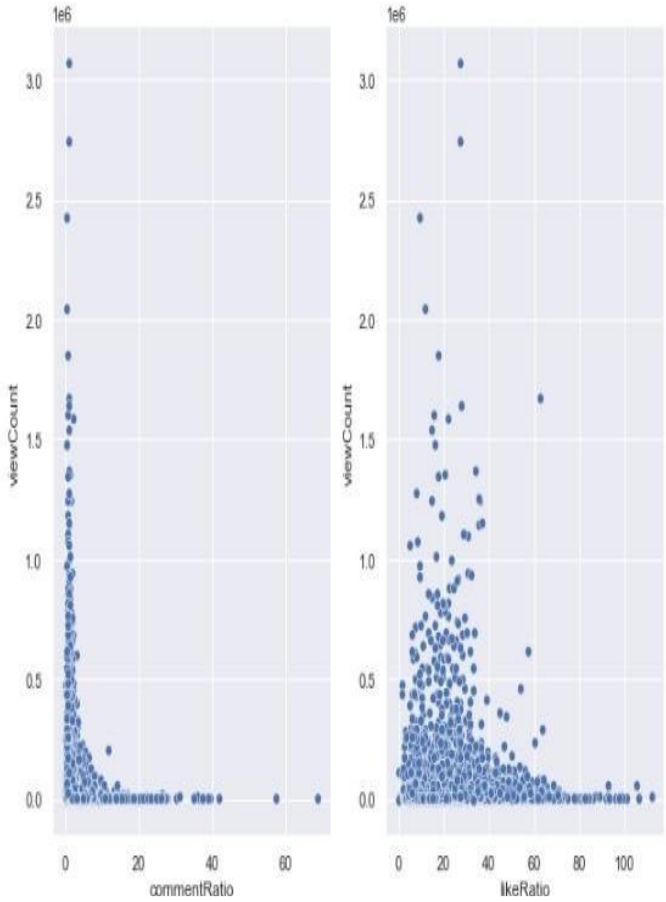


Fig 5: Comment Ratio,like Ratio Vs Viewcount

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 vides 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 relatioship, 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 [ ]: sns.histplot(data=video_df[video_df['durationSecs'] < 10000], x="durationSecs", bins=30)

Out[ ]: <AxesSubplot:xlabel='durationSecs', ylabel='Count'>
```

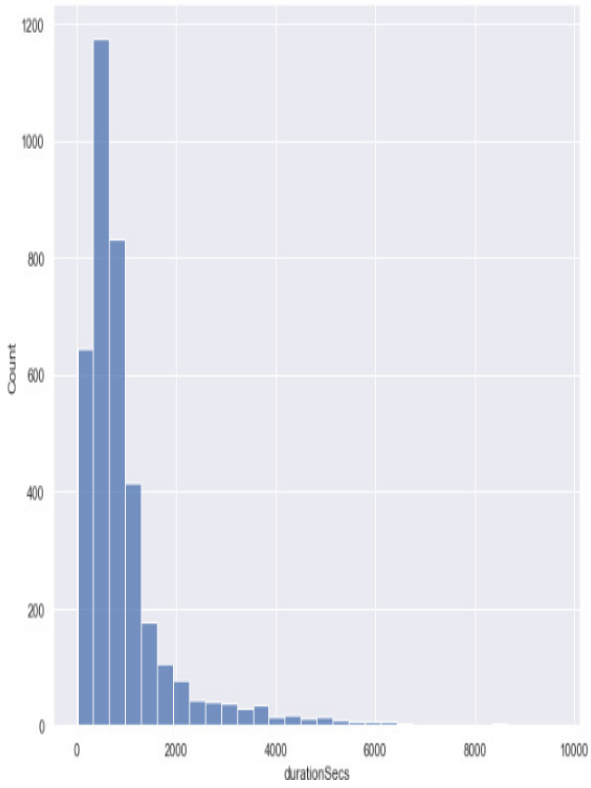


Fig: 6 duration in Secs Vs Count

Does title length matter for views?

There is no clear relationship between title length and views as seen 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'>

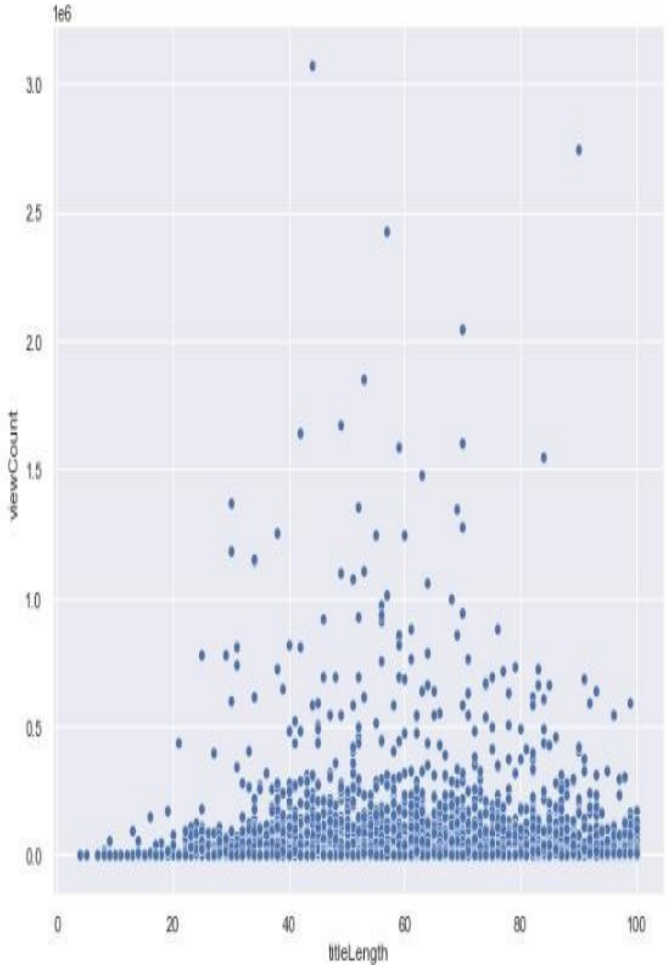
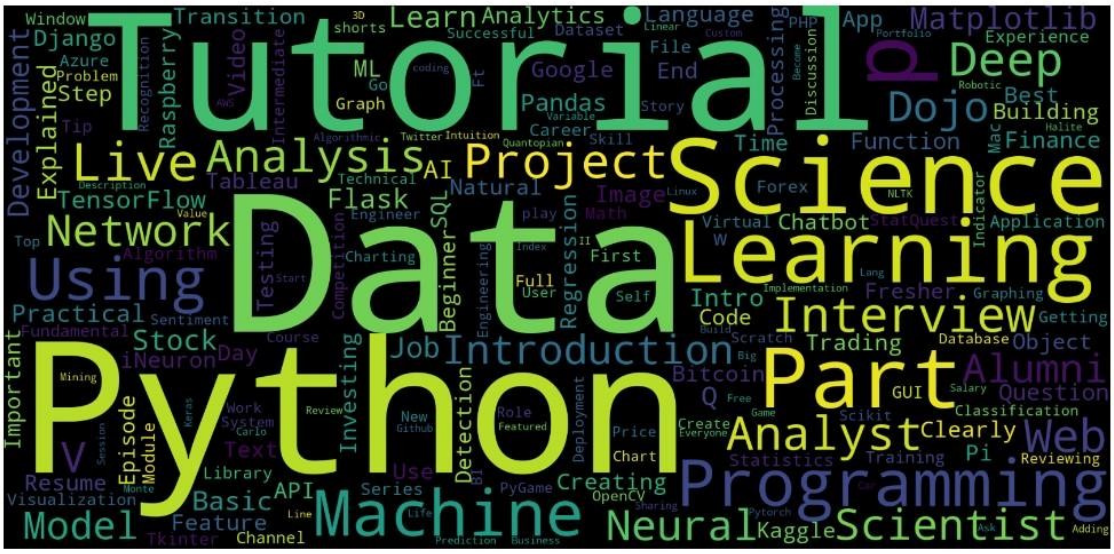


Fig: 7 Titlelength Vs Viewcount



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

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 note contribute a lot to the meaning of the title.

Fig: 9 Word cloud for video comments

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

```
plt.rcParams['figure.figsize'] = (18, 6)
sns.violinplot(video_df['channelTitle'], video_df['viewCount'], palette = 'pastel')
plt.title('Views per channel', fontsize = 14)
plt.show()
```

```
/opt/homebrew/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterp
etation.
  warnings.warn(
```

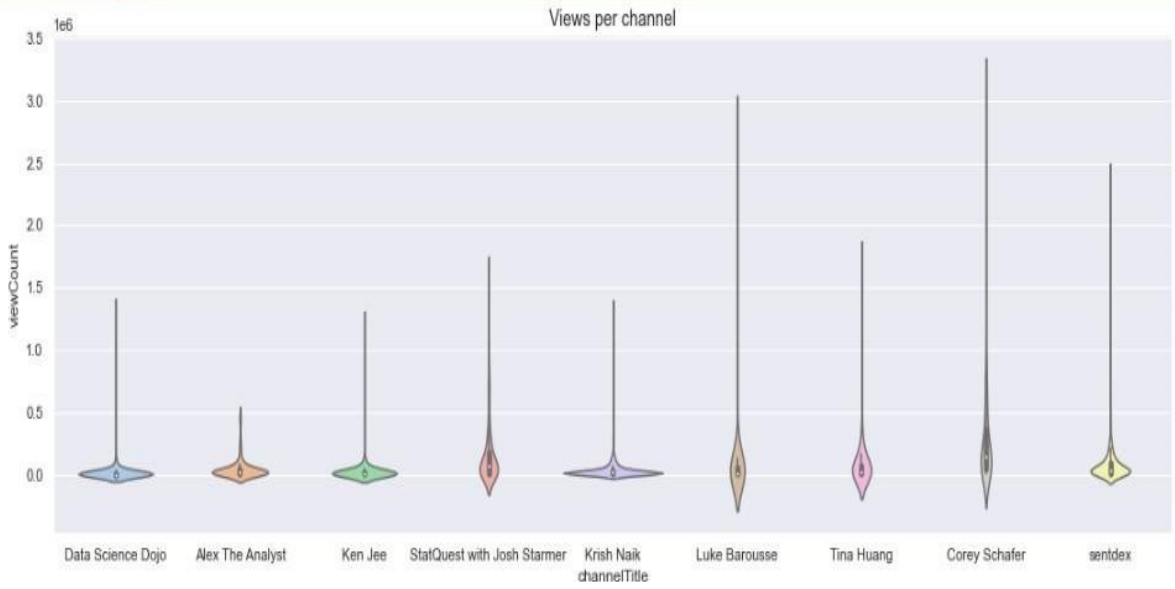


Fig.10 Channel Title Vs Viewcount

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.

```
day_df = pd.DataFrame(video_df['pushblishDayName'].value_counts())
weekdays = [ 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
day_df = day_df.reindex(weekdays)
ax = day_df.reset_index().plot.bar(x='index', y='pushblishDayName', rot=0)
```

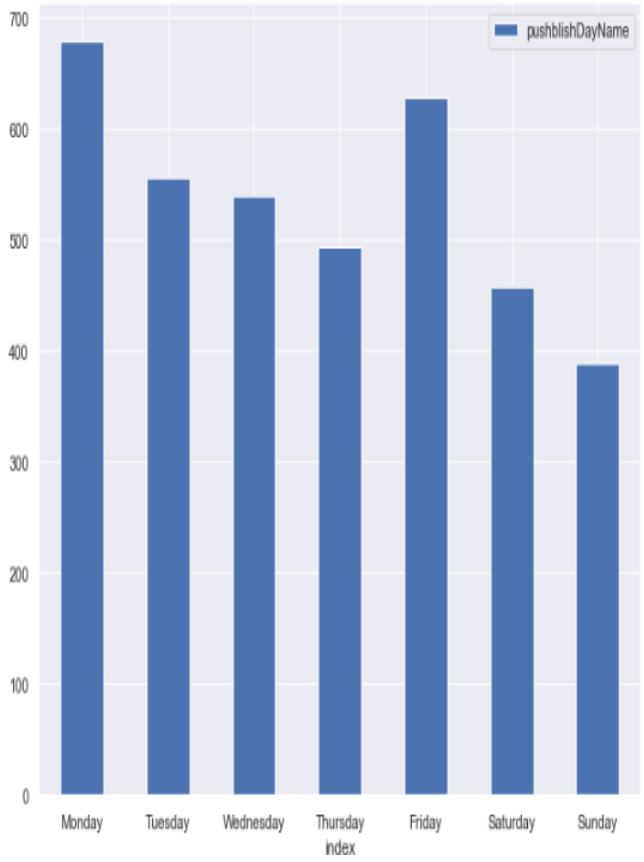


Fig:11 Videos uploaded day wise in a week

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

Enclosure:

1. Copy of Memo from each semester mentioned above.
2. Copy of Schema from each semester mentioned above.

potential market gaps in content that could be filled essential.

References:

[1] Youtube API. Avaiable at https://developers.google.com/youtube/v3

[2] Converting video durations to time function. https://stackoverflow.com/questions/15596753/how-do-i-get-video-durations-with-youtube-api-version-3

[3] P. Covington, J. Adams, E. Sargin. The youtube video recommendation system. In Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '16, pages 191-198, New York, NY, USA, 2016. ACM.Wahyuningsih, Wahyuningsih & Suparman, Suparman & Bachri, Syamsul & Muzakir, Muzakir. (2022). The Marketing Strategy for Tourism Industry Post Covid-19 Pandemic. 10.2991/assehr.k.220707.001.