DATA MINING PRINCIPLES
MSCA 31008

Final Project Submission Youtube Analysis

Austin Sky-Kappler
Carrie Wang

THE UNIVERSITY OF
CHICAGO

Agenda

1) Executive Summary

2) Data
   - Explain
   - Analyze
   - Visualize

1) Data Prep
   - Cleaning / Transformation
   - Resulting Data Variables

1) Modeling
   - Model Fitting
   - Model Performance
   - Model Predictions
   - Feature Relevance

1) Conclusion
   - Findings
   - Next Steps

THE UNIVERSITY OF
CHICAGO

Executive summary

Executive Summary

Problem Statement: As Youtube continues to grow in popularity, content creation on the platform has become mainstream and can be a viable source of income for the creators that operate on it. One issue that these channels may run into is trying to determine what subjects to create content on and then how to structure the video and structure description in a manner that optimizes engagement for the channel.

Project Overview: By connecting to the Youtube API and generating a sample dataset of videos and channels, the team mined and analyzed features to create a model that will predict a video's weighted engagement score (Likes + 2*Comments).
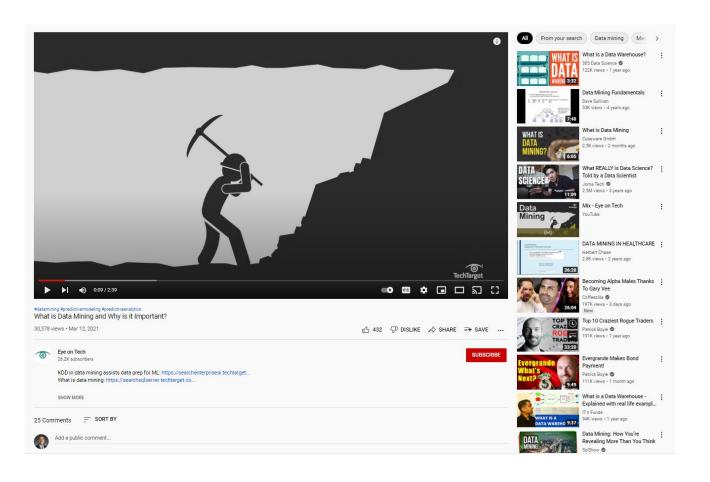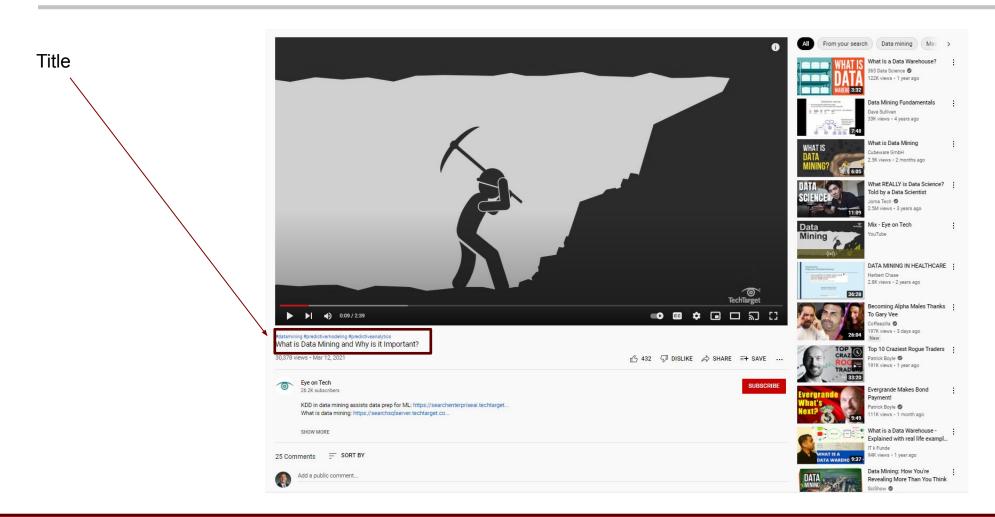
Data Overview - Video - 15,000 Videos Analyzed

Feature	Description	Data Type
video_id	Unique identifier for each video	String
video_title	Title for the video	String
video_description_short	Short description of the video that displays when it is pulled up	String
video_description_full	Displays when user clicks show more	String
video_thumbnail	Image that displays for the video	Object (jpg)
video_upload_date	Date of video upload	
video_view_count	Count of views by users	Int64
video_like_count	Count of likes by users	Int64
video_dislike_count	Count of dislikes by users	Int64
video_comment_count	Count of comments left by users	int64

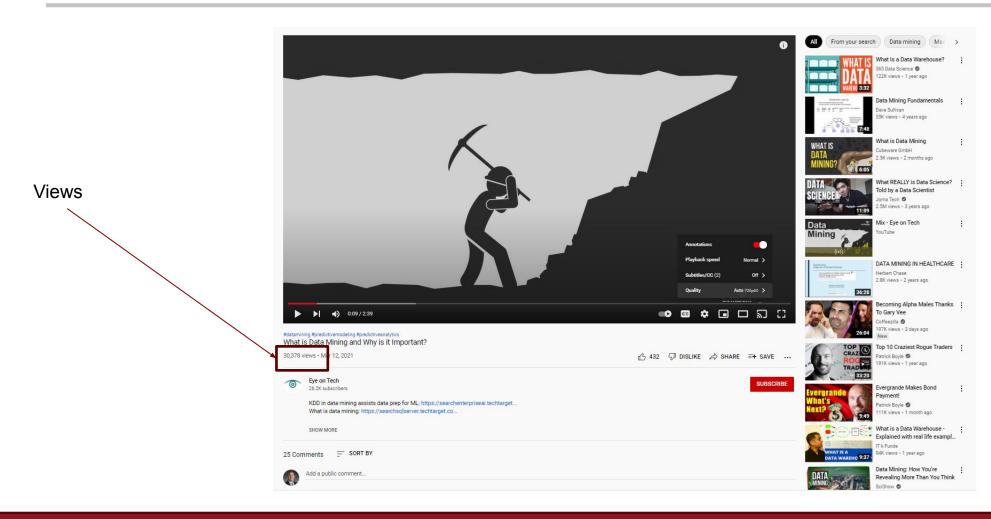| Feature	Description	Data Type
video_tags	Meta tags to support SEO	Object
video_categoryID	ID determining video category	Int64
video_definition	Video quality – SD vs HD	String
video_duration	Length of video	Object
video_caption	Indicates whether captions are available for the video.	Boolean
video_licensedContent	Indicates whether the video represents licensed content, which means that the content was uploaded to a channel linked to a YouTube content partner and then claimed by that partner.	Boolean

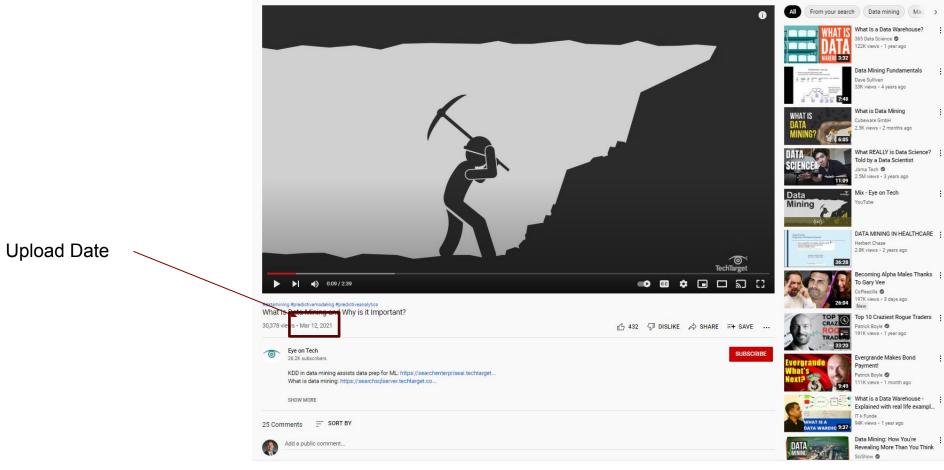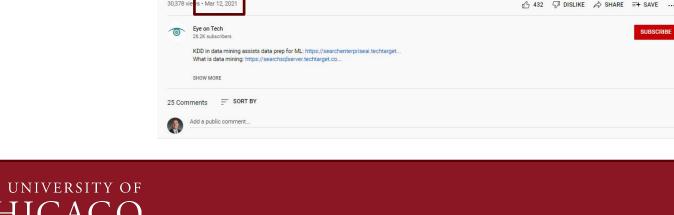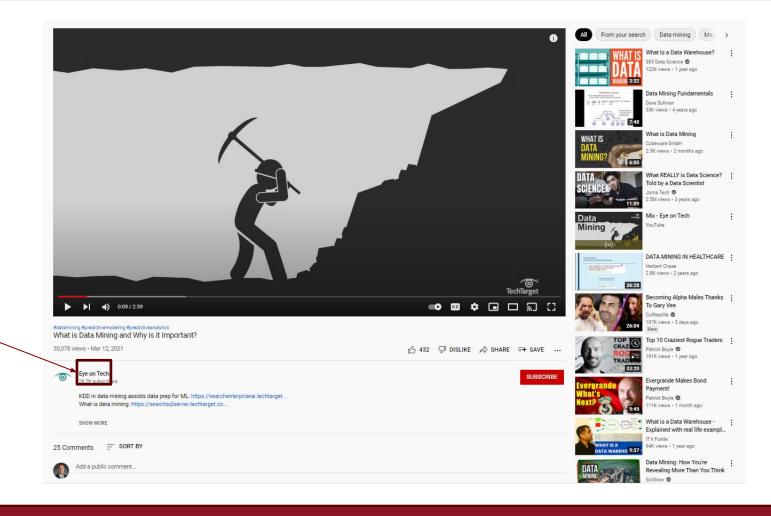Extracted from Youtube.com via Google Dev Platform API Connection from 12/1/21 - 12/7/21

Video
Data
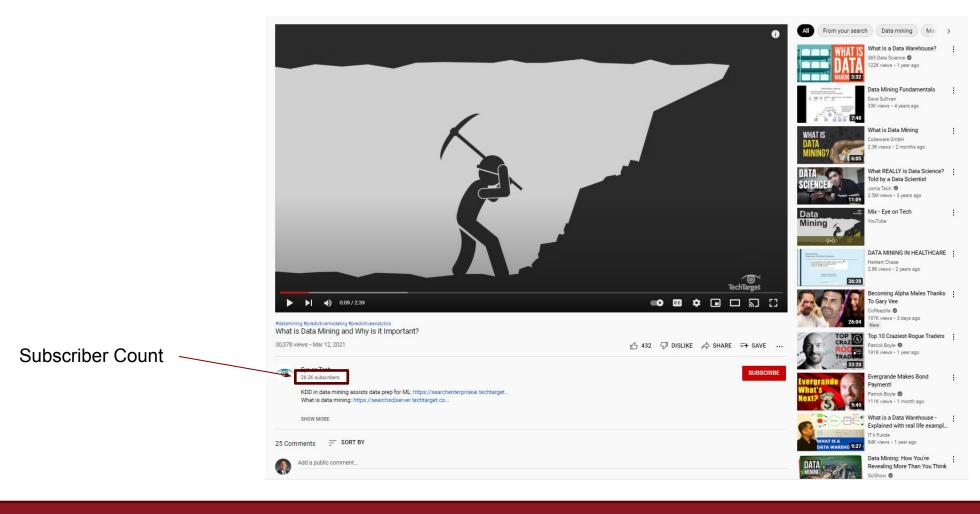# Overview

Video Data Overview

Title

Video Data Overview

Views

Video Data Overview

Upload Date

Video Data Overview

Channel Name

Video Data Overview

Subscriber Count

Video Data Overview

Short Description

Video Data Overview

Long Description

Video Data Overview

Comments

Video Data Overview

Thumbnail

Video Data Overview

Likes / Dislikes

Video Data Overview

Quality

Data Overview - Channel - 200 Channels Analyzed

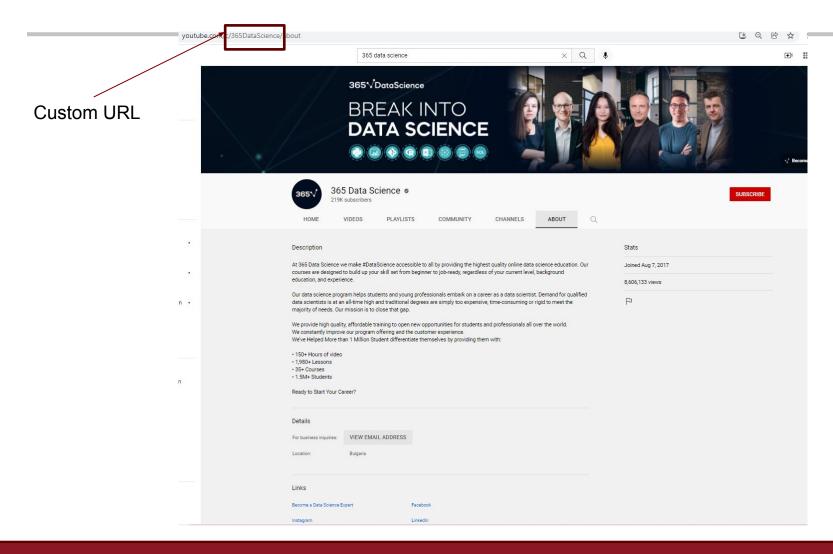| Feature	Description	Data Type
channel_title	Title for Channel	String
channel_description	Description of channel / content	String
channel_customUrl	Creator defined URL vs Youtube generated ID	String
channel_publishedAt	Date Channel first published	Datetime
channel_country	Country of origin for channel	String
channel_subscriberCount	Count of users that elected to receive notifications for the channel	Int64
channel_hiddenSubscriberCount	T/F creator elected to suppress count of subscribers	
channel_viewCount	Total views that all videos on the channel have received	Int64
channel_topicCategories	A list of Wikipedia URLs that describe the channel's content.	Int64
channel_privacyStatus	T/F is the video available for public access	

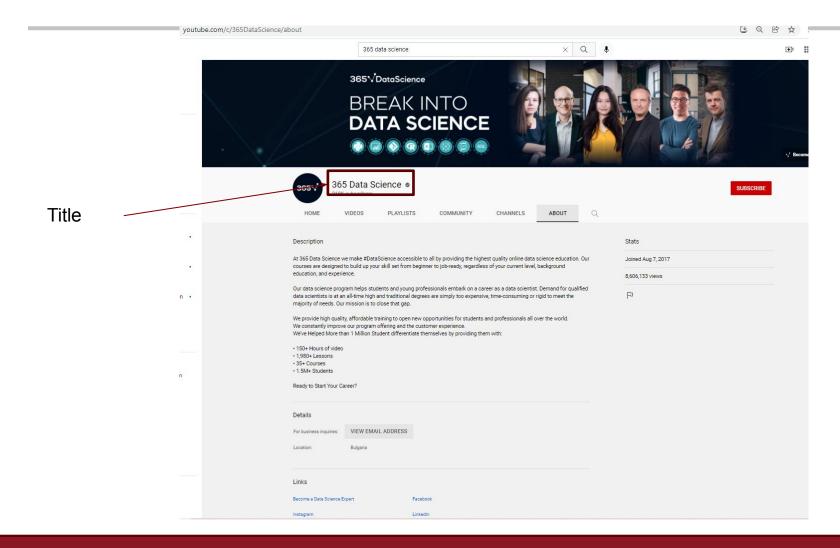| Feature	Description	Data Type
channel_longUploadsStatus	If the channel is eligible to post videos that are > 15 minutes	String
channel_madeForKids	whether the channel is designated as child-directed	Boolean
channel_keywords	Meta tags to support SEO	Int64
channel_videoCount	Total count of videos that have been uploaded by the channel	Int64

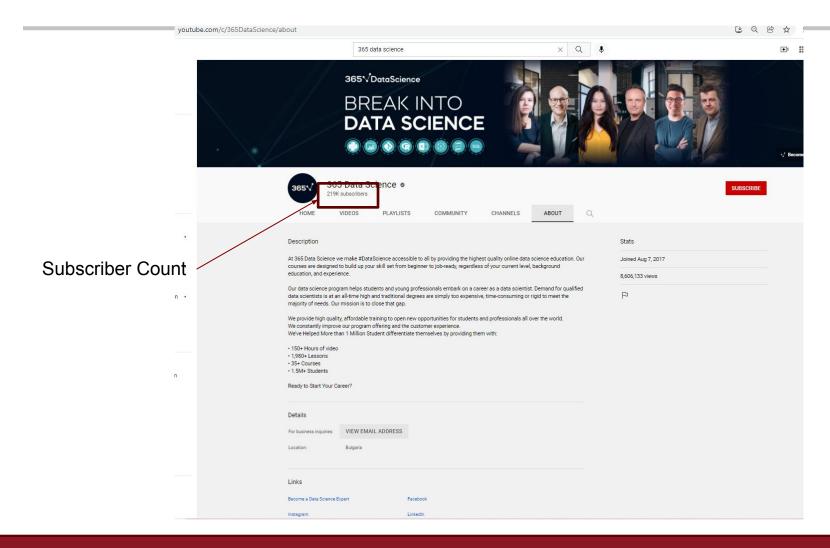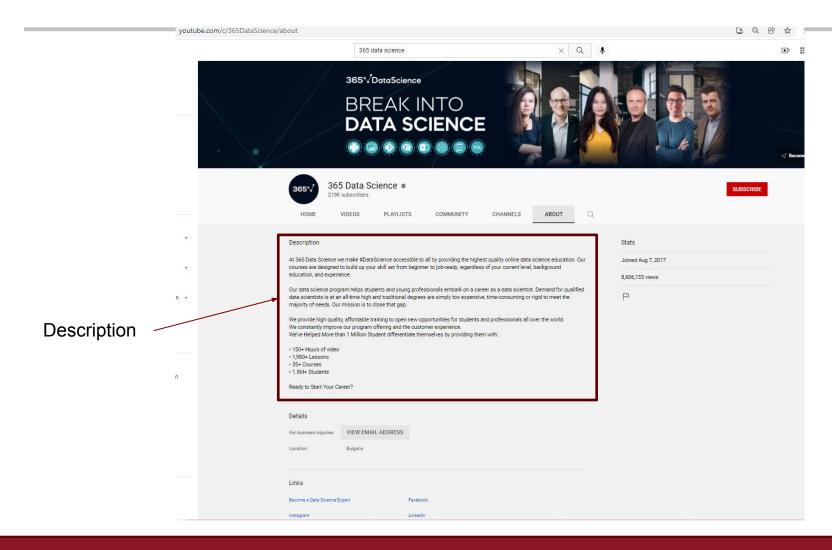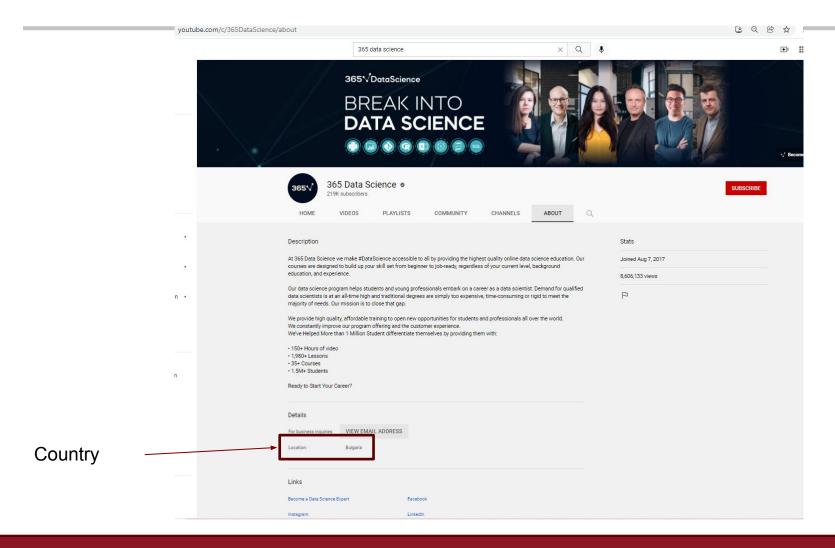Extracted from Youtube.com via Google Dev Platform API Connection from 12/1/21 - 12/7/21

Channel Data Overview
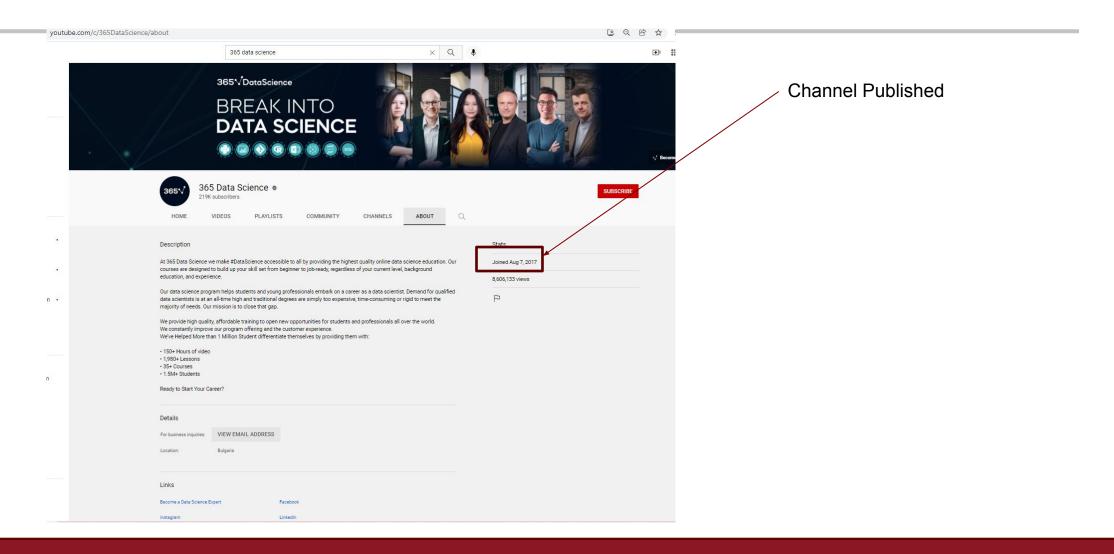
Channel Data Overview

Channel Data Overview

Custom URL

Channel Data Overview

Title

Channel Data Overview

Subscriber Count

Channel Data Overview

Description

Channel Data Overview

Country

Channel Data Overview

Channel Published

Channel Data Overview

Views

Meta Data

Video Page Tags

rue,"kevlar_home_skeleton":true,"kevlar_home_skeleton_hide_later":true,"kevlar_include_query_in_search_endpoint":true,"kevlar_injector":true,"kevlar_inlined_html_templates_polymer_flags":true,"kevlar_js_fixes":true,"kevlar_k

ttp..."><meta name="keywords" content="joma, vlog, data science, data scientist, facebook, google, netflix, amazon, data analyst, business intelligence, AI, deep learning, big data"><link rel="shortlinkUrl" href="https://you
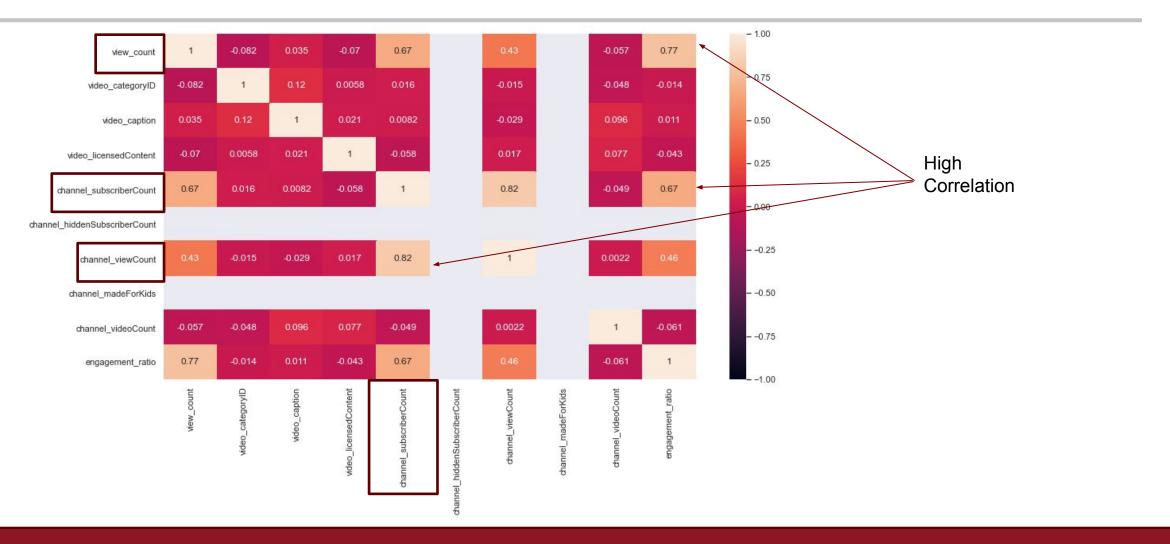
Channel Keyword Tags

e\u003dtrue\u0026preskip_button_style_ads_backend\u003dcountdown_next_to_thumbnail\u0026qoe_nwl_downloads\u003dtrue\u0026qoe_send_and_write\u003dtrue\u0026record_app_crashed_web\u003dtrue\u0026render_unicode_emojis_as_images\u003dtr

)_f9CZUw","keywords":"\"data analyst\" \"data analyst jobs\" \"data analyst\" analytics \"data analyst salary\" \"how to become a data analyst\" \"data analyst career\" sql \"analytics career\" \"data scientist\" \"how to become a d

Feature Heatmap

Data Prep

Data Exploration

```
RangeIndex: 9849 entries, 0 to 9848
Data columns (total 30 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   video_id                    9849 non-null   object
 1   video_title                 9848 non-null   object
 2   video_description_short     9404 non-null   object
 3   video_description_full      9406 non-null   object
 4   video_thumbnail             9849 non-null   object
 5   upload_date                 9849 non-null   object
 6   view_count                  9849 non-null   int64
 7   like_count                  9849 non-null   int64
 8   dislike_count               9849 non-null   int64
 9   comment_count               9823 non-null   float64
 10  video_tags                  8160 non-null   object
 11  video_categoryID            9849 non-null   int64
 12  video_definition            9849 non-null   object
 13  video_duration              9849 non-null   object
 14  video_caption               9849 non-null   bool
 15  video_licensedContent       9849 non-null   bool
 16  channel_title               9849 non-null   object
 17  channel_description         8678 non-null   object
 18  channel_customUrl           8502 non-null   object
 19  channel_publishedAt         9849 non-null   object
 20  channel_country             9406 non-null   object
 21  channel_subscriberCount     9849 non-null   int64
 22  channel_hiddenSubscriberCount 9849 non-null bool
 23  channel_viewCount           9849 non-null   int64
 24  channel_topicCategories     9849 non-null   object
 25  channel_privacyStatus       9849 non-null   object
 26  channel_longUploadsStatus   9849 non-null   object
 27  channel_madeForKids         7731 non-null   float64
 28  channel_keywords            7574 non-null   object
 29  channel_videoCount          9849 non-null   int64
dtypes: bool(3), float64(2), int64(7), object(18)
```
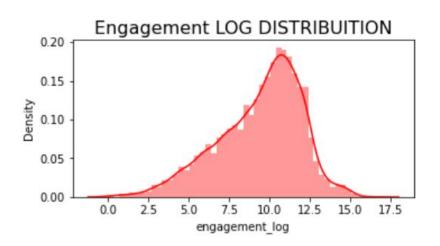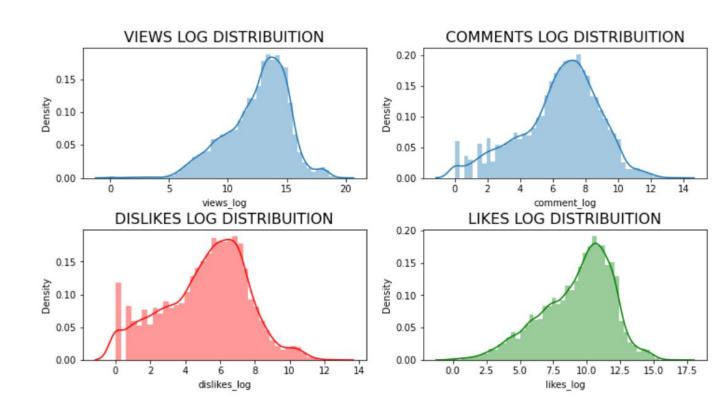
Removed unnecessary columns

Data Exploration

Target: Engagement!

Engagement = 2 * comment + likes
(we weight comments higher)

Engagement LOG DISTRIBUITION

VIEWS LOG DISTRIBUITION

COMMENTS LOG DISTRIBUITION

DISLIKES LOG DISTRIBUITION

LIKES LOG DISTRIBUITION

normal distribution?

Data Exploration

Categorical & Dummy Variables

- Categories?
- Is HD?
- Is in US?
- Is for kids?
- Dummy variables transformed from other numerical variables, such as:
- Is video length between 20 and 60 minutes?
- Is the video new or old?

Numerical Variable

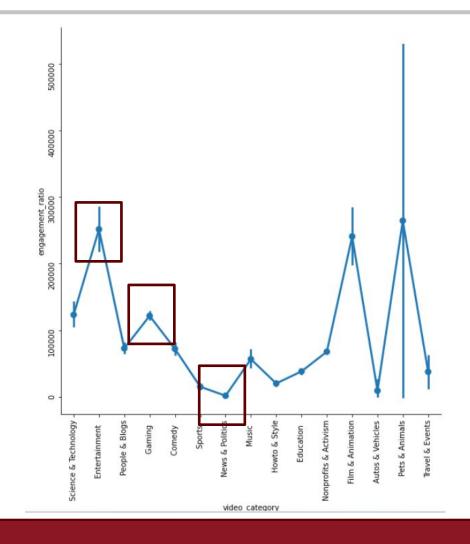- Video duration: how many minutes
- Video upload dates
- Subscribers

Text Analysis

- Video title
- Video description
- Video tags
- Channel description
- Channel keywords

"Engagement"

How to increase engagement from audience?

Data Exploration

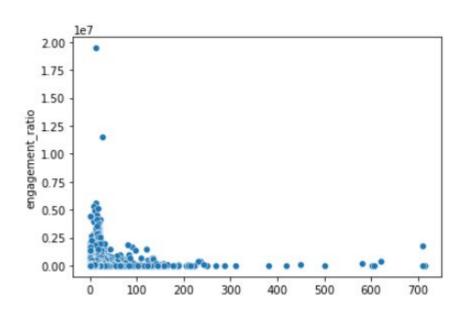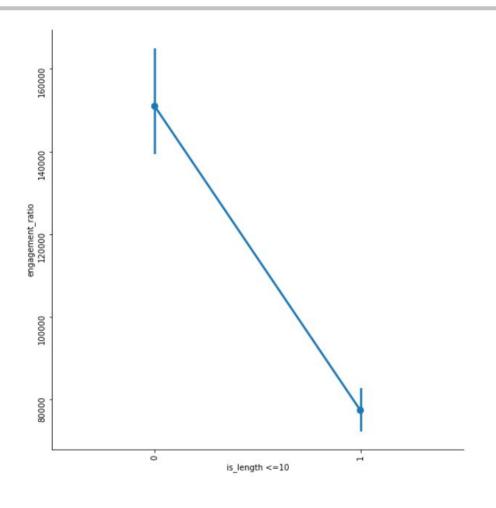# Data Exploration


```
hd      13194
sd        139
Name: video_definition, dtype: int64
```

Data – Video Length

Feature Mining

Categorical & Dummy Variables
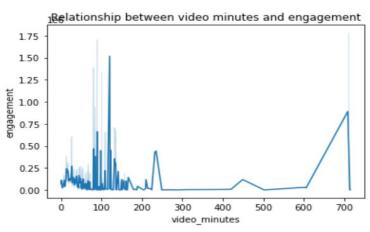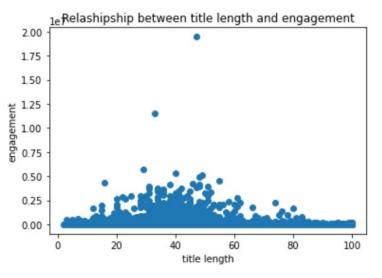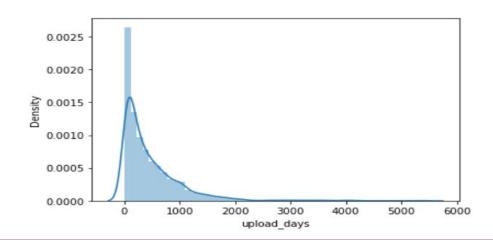
Feature Mining

Numerical Variables

- **Views**
- **Subscribers**
- **Publish time:**
  - frequency of uploading videos every month
  - how many days after being uploaded

Feature Mining

Text Analysis

Text Pre-processing
- Remove non-words(characters, symbols, etc.)
- Extract all keywords from description using NLP

Text Encoding
- Transform words into vectorizers using **TF-IDF** Vectorizer, so similar words are mapped near each other

Clustering
- Elbow Method
- **Kmeans:** gave 9 clusters
- Get labels for every video based on the word similarity
- Tried **DBSCAN**, gave 2 labels for all data

Feature Mining

Text Analysis

Is TF-IDF a good encoding method for clustering?

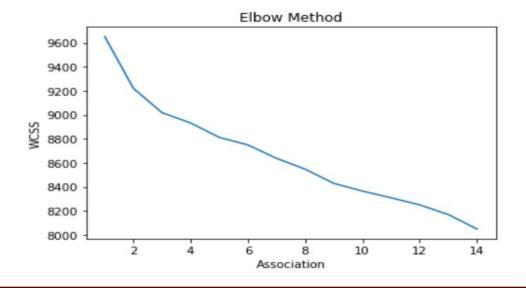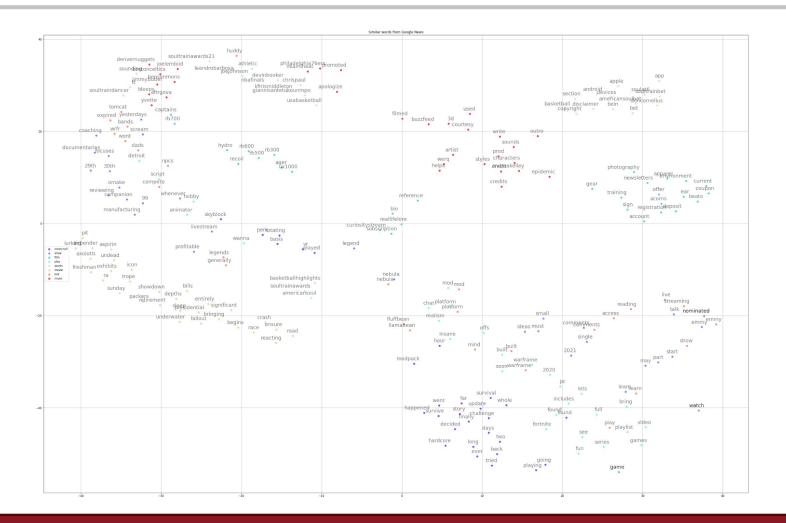After using TF-IDF and getting all the words, we found that those similar words cluster because they look similar alphabetically, not because they have the same meaning.

```
'admin', 'administer', 'administration', 'administrative', 'admins', 'admire', 'admiring
mitting', 'ado', 'adobe', 'adobo', 'adoboli', 'adolescence', 'adolf', 'adolfo', 'adolph'
ing', 'adoption', 'adoptive', 'adopts', 'adorable', 'adorablebirbpon', 'adorably', 'ado
riangranados', 'adrien', 'ads', 'adult', 'adulthood', 'adulting', 'adults', 'advance',
s', 'advancing', 'advantage', 'advantaged', 'advantages', 'advent', 'adventure', 'adven
'adversaries', 'adversary', 'adverse', 'adversity', 'advertise', 'advertised', 'adverti
ng', 'advice', 'advised', 'adviser', 'advising', 'advisor', 'advisors', 'advocacy', 'ad
e', 'aea', 'aeneid', 'aerial', 'aero', 'aerobatic', 'aerobatics', 'aerodynamic', 'aerod
spyke', 'aeryn', 'aeschliman', 'aesthetic', 'aesthetically', 'aesthetics', 'aesuno', 'a
dark', 'aewdarkelevation', 'aewdynamite', 'aewfullgear', 'aewrampage', 'af', 'afaik', '
ct', 'affected', 'affectent', 'affecting', 'affection', 'affectionate', 'affects', 'aff
ffinity', 'affion', 'affioncrocket', 'affirming', 'affleck', 'afflict', 'affluence', 'a
```

Feature Mining

Text Analysis

Word2Vec and TSNE

We set key words as labels manually, like "minecraft", "show", "film", "play", "sports", "movie", "talk", "music" . We let similar videos cluster around those keys.
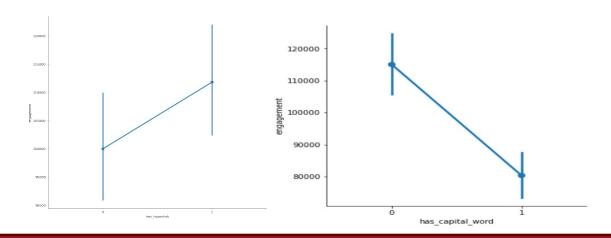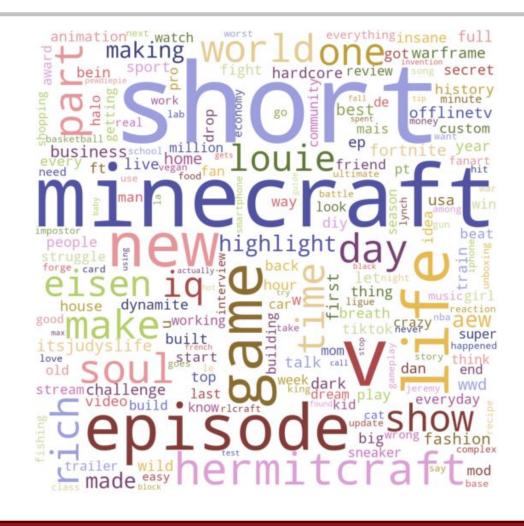
Text Analysis

Generate more values from text

- For "Video Title", calculated the title length, found that most videos have titles between 30-60 words
- Found most common words appearing in all the video titles
  - video description contains hyperlink?
  - video title contains CAPITAL WORDS, like "HOW"?

Modeling

Model Performance

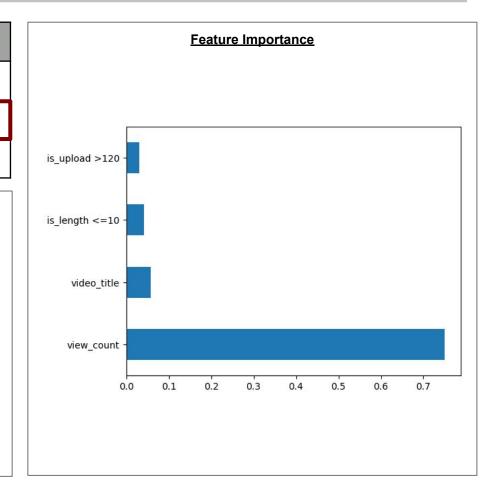| Model	Train RMSE	Test RMSE	Cross Validation	
Random Forest	84,103	104,971	236,226	.89
Linear Regression	225,549	151,184	226,017	.77
Decision Tree	189,455	112,099	216,817	.87

While Random Forest had the highest accuracy, the current apparent over fitting leads the team to select Linear Regression the purpose of this exercise

As additional channels/videos were added, RMSE, prediction, and over-fitting improved

Additional models that were tested were Gradiant Boosting and Bagging Classifier

Models were ran on 3 different sets of features while optimizing

The team believes that the low Test RMSE associated with test data was related to the sample size and characteristics of this dataset
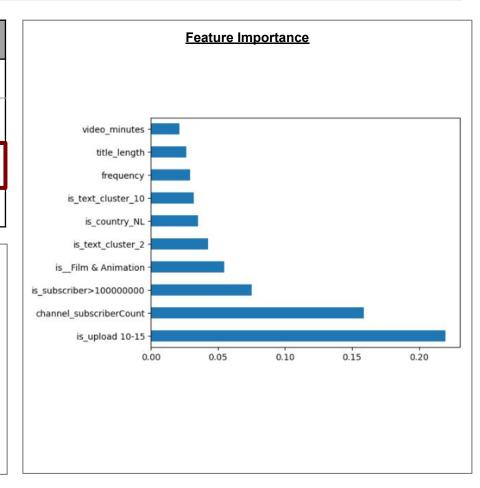
Feature Importance

Model Performance

Model	Train RMSE	Test RMSE	Cross Validation	
Random Forest	97449.11	97449.11	244221	.67
Linear Regression	241899	152735	244441	.61
Decision Tree	229753	136638	232276	.69
XGradiant Boost	69350	157254	295522	.67

Based on feedback received on Tuesday, following the presentation, the team attempted to re-run the model by using word2vec, and without using view_count as a predictor

With enhancements and modifications, the team was able to still achieve ~.7 accuracy

The updated model is optimized when using Decision Tree, but still suffers from some level of overfitting and would benefit from a larger sample of channels and videos

Feature Importance

Conclusion

- **The higher the views, the higher the engagement**

- **Don't make your video too long! Or too short!**

Feature Mining:

- KMeans: the number of clusters needs to be determined in advance. When the data set is relatively large, it is difficult to give an appropriate value in advance

Modeling:

- The current model is heavily reliant on the linear relationship with number of views to determine engagement

Next Steps

- Generate more data!

- Continued feature refinement

  - Text Clustering: Word2Vec + TSNE

- Experimenting with building views into the predicted variable

  (likes + comments) / views

THE UNIVERSITY OF
CHICAGO