MA 678 Final Project

Handing Zhang

12/7/2021

Abstract

I conducted a multilevel linear regression model to find the relationship between the count of likes and certain subset of features of the videos. I used the category of videos as my groups for random effect evaluation. **Research Question:** factors that contribute to the number of likes. **Random Effect** Categories of video. **Fixed effects:** video_age, duration_sec, caption

Introduction:

bbc has a youtube channel where it posts different kinds of videos everyday. Some of the videos receive a lot of likes from viewers while the others not so much. An interesting topic ,then, to study is that what are the factors that influence the number of likes.

The dataset I use is published on Kaggle: name: ***BBC YouTube Videos Metadata*** link: https://www.kaggle.com/gpreda/bbc-youtube-videos-metadata

column names	explanation
video_title	The title of the video
days_since_published	number of days from date of publish to 2021-12-07
category	The category of the video
duration_sec	How long is the video in seconds
view_count	The number of views
like_count	The number of likes of the video
dislike_count	The number of dislikes of the video
caption	Boolean value indicating whether or not there is caption
comment_count	number of comments

Data Cleaning

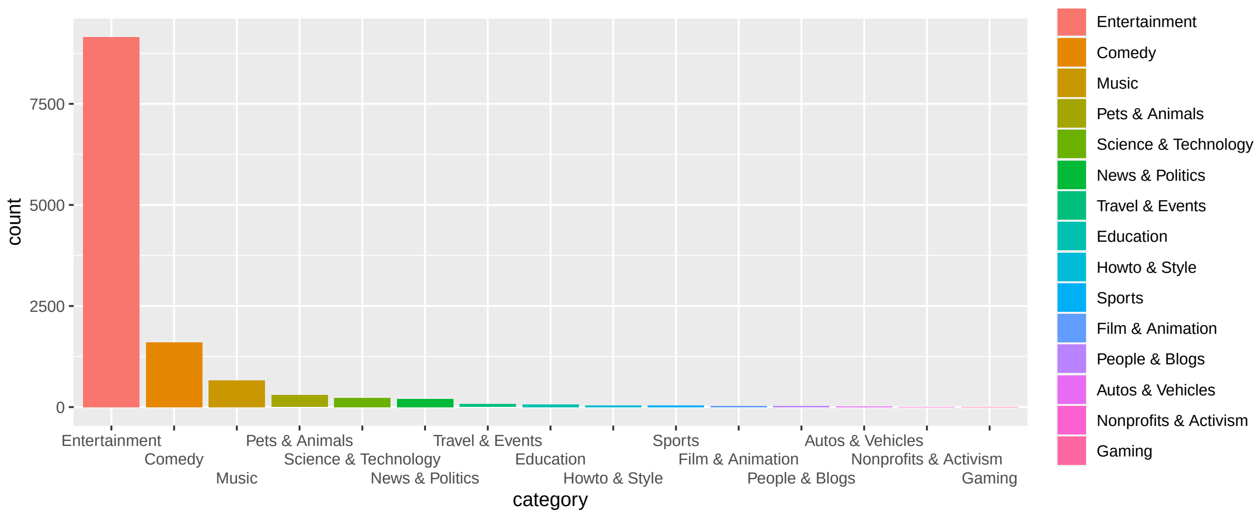
First I performed some data wrangling after reading in the data. I created a new column called "days_since_published", which is the number of days from date of publish to 2021-12-07.

I noticed that there were some NAs in numeric columns. I chose to conducted a multiple imputation on the missing values.

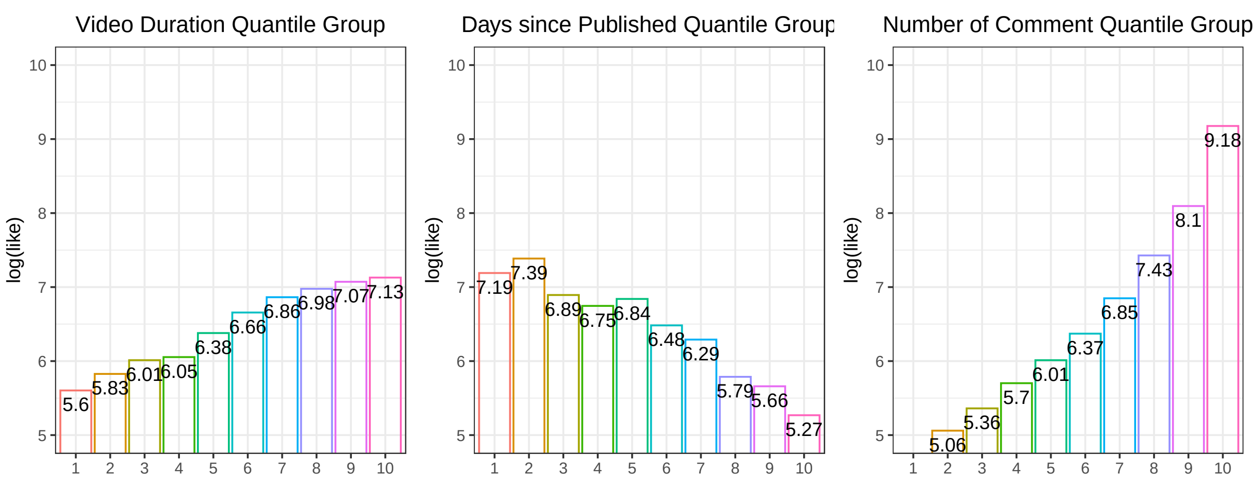
Therefore, I performed a multiple imputation on the missing values of bbc.

I took natural logarithm of several variables: number of comments, number of views, number of likes, number of dislikes and days since published.

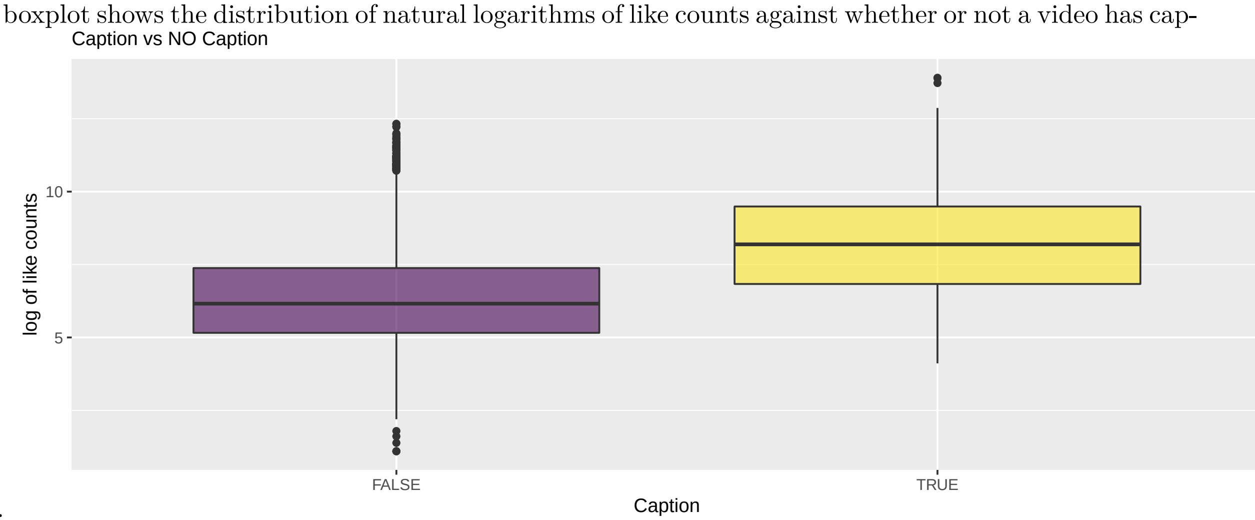
EDA



The two plots below shows the natural logarithms of like counts grouped by 10 quantiles of video duration and days since published.

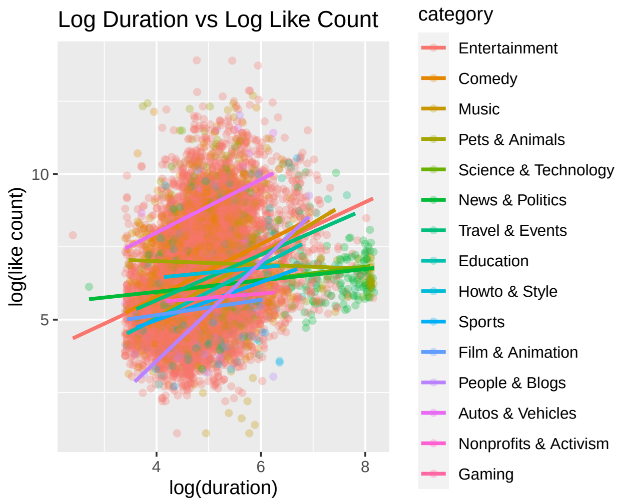


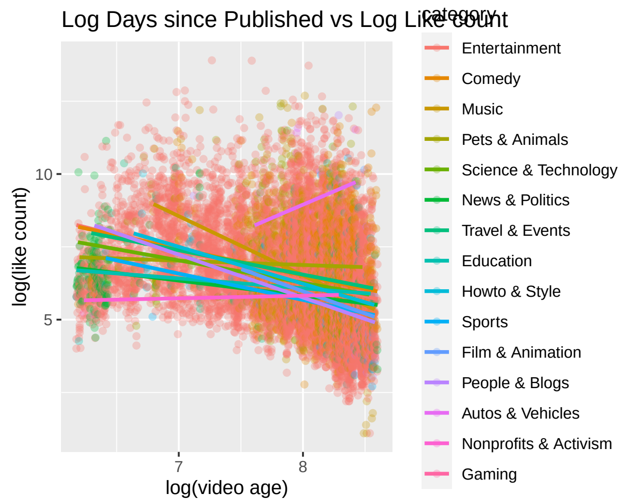
The boxplot shows the distribution of natural logarithms of like counts against whether or not a video has cap-



tion.

The following two plots shows a general relationship between duration, days since published and number of likes received by a video.





We can see from the plot that in most categories there exists a postive relationship between the duration of a video and the number of the likes it receives. On the other hand, there are usually a negative relationship between days since published and the number of likes, with Autos and Vehicles videos as an exception. It seems the older the video is, the less likes it receives.

Method:

Model Fitting

I fitted a multilevel model with duration, days since published and caption as my fixed effects, where I combined duration and random effects in my model.

** fit_2 <- lmer(log_like ~ log_duration + log_age + caption + (1 + log_duration|category), data = bbc) **

Result: What you found.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	9.282e+00	7.161e-01	1.642e+01	12.962	4.79e-10 ***
log_duration	4.075e-01	1.279e-01	1.335e+01	3.186	0.00695 **
log_age	-6.349e-01	2.925e-02	1.240e+04	-21.709	< 2e-16 ***
CaptionTRUE	1.121e+00	6.302e-02	1.242e+04	17.785	< 2e-16 ***

We can see the fixed effects below, all variables are significant at alpha = 0.05 level.

For Entertainment videos as an example for our group category:

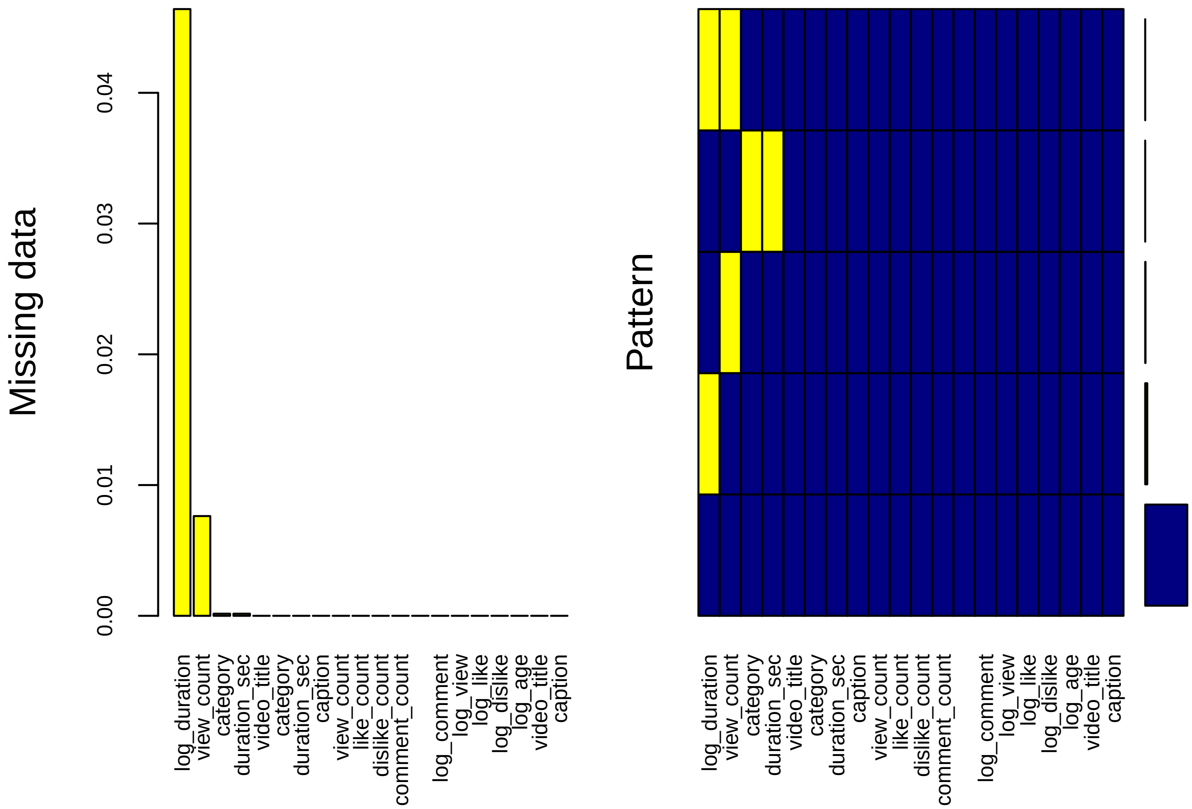
$$y = 8.730573 + 0.53904341\beta_{log-duration} - 0.6348909\beta_{log-age} + 1.120828\beta_{Caption}$$

Discussion:

The model demonstrates that in general long duration having caption have a positive impact on the average number of likes a video receives when fixing other factors. On the other hand the age of a video has a negative effect on the average number of likes when other components stay the same. Next step: I should conduct more model validations to optimize my model

Appendix

Missing Value in Data before Multiple Imputation




```
##
##  Variables sorted by number of missings:
##             Variable        Count
##         log_duration 0.0464033398
##           view_count 0.0076268465
##             category 0.0001605652
##         duration_sec 0.0001605652
##          video_title 0.0000000000
##             category 0.0000000000
##         duration_sec 0.0000000000
##              caption 0.0000000000
##           view_count 0.0000000000
##           like_count 0.0000000000
##        dislike_count 0.0000000000
##        comment_count 0.0000000000
##  days_since_published 0.0000000000
```

```
##           log_comment 0.0000000000
##              log_view 0.0000000000
##              log_like 0.0000000000
##           log_dislike 0.0000000000
##               log_age 0.0000000000
##           video_title 0.0000000000
##               caption 0.0000000000
```

Detail of Model Fitted

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: log_like ~ log_duration + log_age + caption + (1 + log_duration |
##     category)
##    Data: bbc
##
## REML criterion at convergence: 46126
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5433 -0.7220 -0.1332  0.6197  3.9561
##
## Random effects:
##  Groups   Name         Variance Std.Dev. Corr
##  category (Intercept)  4.2412   2.0594
##           log_duration 0.1573   0.3966   -0.94
##  Residual              2.3734   1.5406
## Number of obs: 12437, groups:  category, 15
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)   9.282e+00  7.161e-01  1.642e+01  12.962 4.79e-10 ***
## log_duration  4.075e-01  1.279e-01  1.335e+01   3.186  0.00695 **
## log_age      -6.349e-01  2.925e-02  1.240e+04 -21.709  < 2e-16 ***
## captionTRUE   1.121e+00  6.302e-02  1.242e+04  17.785  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) lg_drt log_ag
## log_duratin -0.901
## log_age     -0.370  0.046
## captionTRUE -0.111 -0.008  0.345
```

```
## $category
##                       (Intercept) log_duration    log_age captionTRUE
## Entertainment            8.730573   0.53904341 -0.6348909    1.120828
## Comedy                   7.808430   0.79930504 -0.6348909    1.120828
## Music                    7.012601   0.85676514 -0.6348909    1.120828
## Pets & Animals          13.197388  -0.28521856 -0.6348909    1.120828
## Science & Technology    10.426876   0.15278389 -0.6348909    1.120828
## News & Politics         11.047085  -0.04807524 -0.6348909    1.120828
## Travel & Events          9.221300   0.45651868 -0.6348909    1.120828
```

```
## Education                8.647776   0.46177589 -0.6348909    1.120828
## Howto & Style           10.250364   0.21904199 -0.6348909    1.120828
## Sports                   8.572852   0.43435062 -0.6348909    1.120828
## Film & Animation         9.002807   0.35407126 -0.6348909    1.120828
## People & Blogs           6.562501   0.81264941 -0.6348909    1.120828
## Autos & Vehicles        10.107145   0.64209847 -0.6348909    1.120828
## Nonprofits & Activism    9.251213   0.30952169 -0.6348909    1.120828
## Gaming                   9.395133   0.40852449 -0.6348909    1.120828
##
## attr(,"class")
## [1] "coef.mer"
```

Model Validation

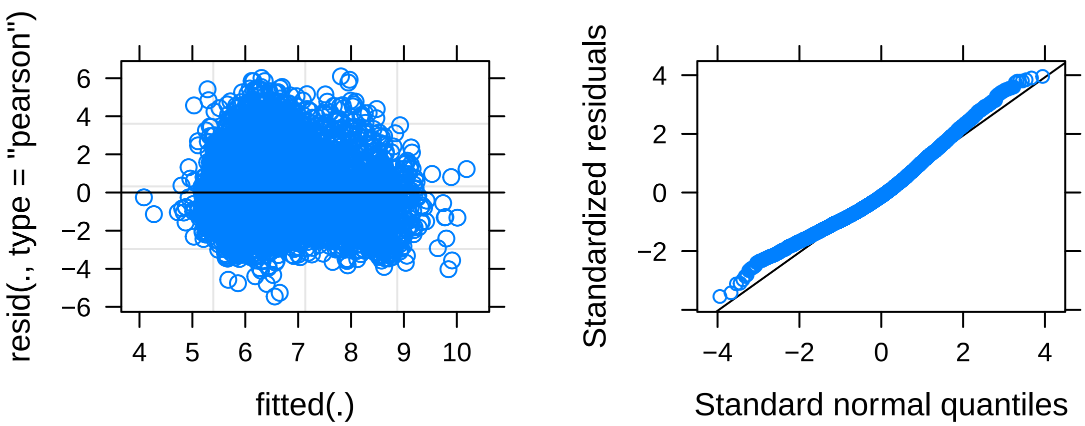


Figure 1: Residual plot and Q-Q plot.

From the qqplot we can see that one limitation of my model is that the residuals do not rigorously follow a normal distribution.

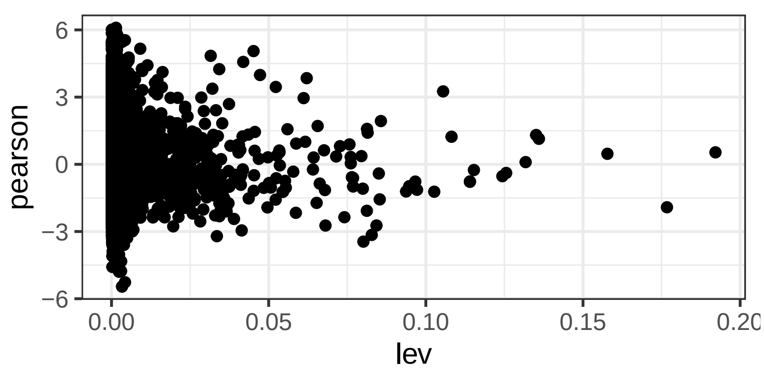
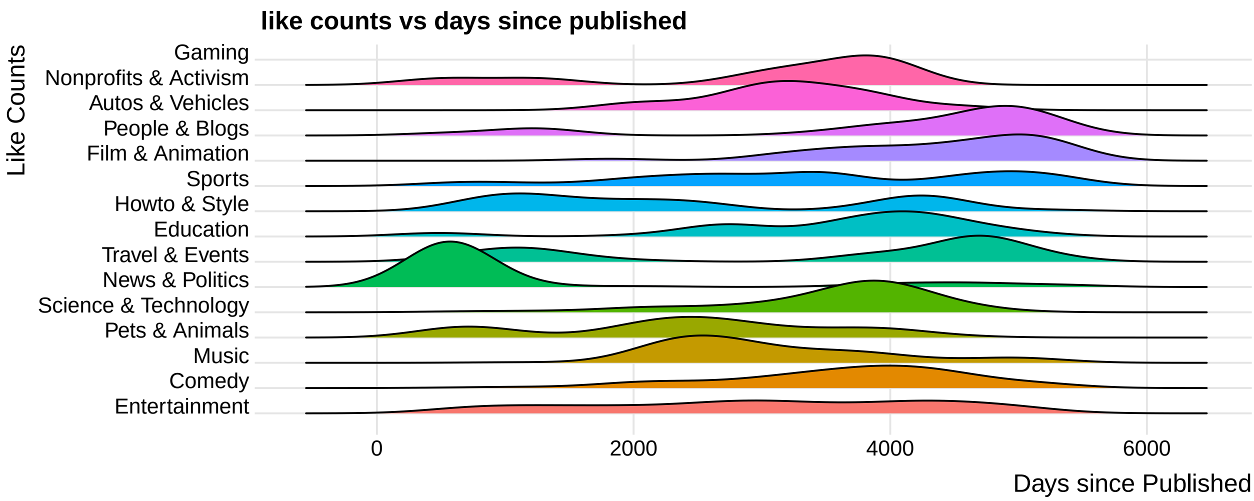
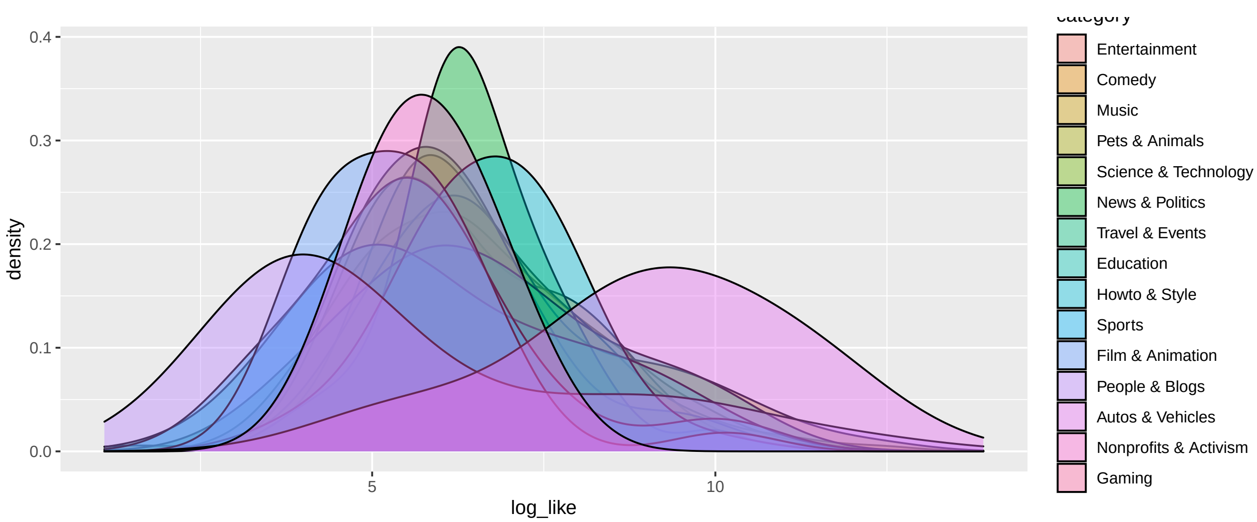


Figure 2: Residuals vs Leverage.

More EDA





Citation

https://github.com/yurijin98/MA678MidtermProject https://www.youtube.com/c/BBCNews https://www.kaggle.com/gpreda/bbc-youtube-videos-metadata