Code ▾

DS606-Project Proposal

George Cruz

2020-11-01

Data Preparation

Hide

```
# load data
library(tidyverse)
library(scales)
library(infer)
library(psych)
library(httr)
library(jsonlite)
```

The Data Set was obtained from Kaggle. This dataset was collected using the YouTube API.

Loading the Data.

Hide

```
#Get the videos csv
raw_video_df <- read_csv(file="https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv",quote = "\"")
```

```
## Parsed with column specification:
## cols(
##   video_id = col_character(),
##   trending_date = col_character(),
##   title = col_character(),
##   channel_title = col_character(),
##   category_id = col_double(),
##   publish_time = col_datetime(format = ""),
##   tags = col_character(),
##   views = col_double(),
##   likes = col_double(),
##   dislikes = col_double(),
##   comment_count = col_double(),
##   thumbnail_link = col_character(),
##   comments_disabled = col_logical(),
##   ratings_disabled = col_logical(),
##   video_error_or_removed = col_logical(),
##   description = col_character()
## )
```

```
## Warning: 1533544 parsing failures.
## row  col           expected actual                                                                      file
##   2 tags delimiter or quote      | 'https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv'
##   2 tags delimiter or quote      l 'https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv'
##   2 tags delimiter or quote      | 'https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv'
##   2 tags delimiter or quote      j 'https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv'
##   2 tags delimiter or quote      | 'https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv'
## ... .... ................. ...... .........................................................................
## See problems(...) for more details.
```

Hide

```
#get the categories JSON
url <- paste("https://raw.githubusercontent.com/georg4re/ds606/main/data/US_category_id.json", sep="")
res <- GET(url)
data <- fromJSON(rawToChar(res$content))

category_df <- data$items %>%
  flatten(.) %>%
  rename(category=snippet.title)
```

Joining the data and the Categories

Hide

```
category_df <- category_df %>%
  rename(category_id = id)
category_df$category_id <- as.numeric(category_df$category_id)


video_df <- raw_video_df %>%
  left_join(category_df) %>%
  select(video_id,
         trending_date,
         title,
         channel_title,
         category,
         publish_time,
         tags,
         views,
         likes,
         dislikes,
         comment_count,
         comments_disabled,
         ratings_disabled,
         video_error_or_removed,
         description
         )
```

```
## Joining, by = "category_id"
```

A snippet

Hide

```
glimpse(video_df)
```

```
## Rows: 40,949
## Columns: 15
## $ video_id               <chr> "2kyS6SvSYSE", "1ZAPwfrtAFY", "5qpjK5DgCt4",...
## $ trending_date          <chr> "17.14.11", "17.14.11", "17.14.11", "17.14.1...
## $ title                  <chr> "WE WANT TO TALK ABOUT OUR MARRIAGE", "The T...
## $ channel_title          <chr> "CaseyNeistat", "LastWeekTonight", "Rudy Man...
## $ category               <chr> "People & Blogs", "Entertainment", "Comedy",...
## $ publish_time           <dttm> 2017-11-13 17:13:01, 2017-11-13 07:30:00, 2...
## $ tags                   <chr> "SHANtell martin", "last week tonight trump ...
## $ views                  <dbl> 748374, 2418783, 3191434, 343168, 2095731, 1...
## $ likes                  <dbl> 57527, 97185, 146033, 10172, 132235, 9763, 1...
## $ dislikes               <dbl> 2966, 6146, 5339, 666, 1989, 511, 2445, 778,...
## $ comment_count          <dbl> 15954, 12703, 8181, 2146, 17518, 1434, 1970,...
## $ comments_disabled      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
## $ ratings_disabled       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
## $ video_error_or_removed <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
## $ description            <chr> "SHANTELL'S CHANNEL - https://www.youtube.co...
```

Hide

```
knitr::kable(head(video_df%>% select(-description),10))
```

video_id	trending_date	title	channel_title	category	publish_time	tags
2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	People & Blogs	2017-11-13 17:13:01	SHANtell martin

video_id	trending_date	title	channel_title	category	publish_time	tags
1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with John Oliver (HBO)	LastWeekTonight	Entertainment	2017-11-13 07:30:00	last week tonight trump presidency“|”last week tonight donal
5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Lele Pons	Rudy Mancuso	Comedy	2017-11-12 19:05:24	racist superman“|”rudy“|”mancuso“|”king“|”bach“|”racist“|”s video“|”iphone x by pineapple“|”lelepons“|”hannahstocking“|”rudymancuso“|”ina My Driver’s License | Lele Pons
puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	Entertainment	2017-11-13 11:00:04	rhett and link“|”gmm“|”good mythical morning“|”rhett and lin morning“|”Season 12“|”nickelback lyrics“|”nickelback lyrics re nickelback“|”gmm nickelback“|”lyrics (website category)“|”nic kroeger“|”canada“|”music (industry)“|”mythical“|”gmm challe
d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	Entertainment	2017-11-12 18:01:41	ryan“|”higa“|”higatv“|”nigahiga“|”i dare you“|”idy“|”rhpc“|”da
gHZ1Qz0KiKM	17.14.11	2 Weeks with iPhone X	iJustine	Science & Technology	2017-11-13 19:07:23	ijustine“|”week with iPhone X“|”iphone x“|”apple“|”iphone“|”i
39idVpFF7NQ	17.14.11	Roy Moore & Jeff Sessions Cold Open - SNL	Saturday Night Live	Entertainment	2017-11-12 05:37:17	SNL“|”Saturday Night Live“|”SNL Season 43“|”Episode 1730“|” Sessions“|”Kate McKinnon“|”s43“|”s43e5“|”episode 5“|”live“|” night“|”host“|”music“|”guest“|”laugh“|”impersonation“|”acto Winfrey“|”OWN“|”Girls Trip“|”The Carmichael Show“|”Keanu“ open
nc99ccSXST0	17.14.11	5 Ice Cream Gadgets put to the Test	CrazyRussianHacker	Science & Technology	2017-11-12 21:50:37	5 Ice Cream Gadgets“|”Ice Cream“|”Cream Sandwich Maker“|” to the Test“|”testing“|”10 Kitchen Gadgets“|”7 Camping Coffe
jr9QtXwC9vc	17.14.11	The Greatest Showman | Official Trailer 2 [HD] | 20th Century FOX	20th Century Fox	Film & Animation	2017-11-13 14:00:23	Trailer“|”Hugh Jackman“|”Michelle Williams“|”Zac Efron“|”Zen school musical“|”hugh jackman musical“|”zac efron musical“|” Barnum“|”Barnum and Bailey“|”Barnum Circus“|”Barnum an trailer“|”the greatest showman trailer“|”logan“|”Benj Pasek“|”
TUmyygCMMGA	17.14.11	Why the rise of the robots won’t mean the end of work	Vox	News & Politics	2017-11-13 13:45:16	vox.com“|”vox“|”explain“|”shift change“|”future of work“|”au shierholz“|”martin ford“|”rise of the robots“|”humans“|”work income

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for. Is it possible to predict based on these variables or a combination of them the popularity of a youtube video in America?

Cases

What are the cases, and how many are there? Each observation represents a video in Youtube. There are 40,949 observations.

Data collection

Describe the method of data collection. Data was obtained from a Kaggle data set. (https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv)

Type of study

What type of study is this (observational/experiment)? This is an observational study based on the obervations captured in this data.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link. Data was obtained from a Kaggle data set. (https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv)

Dependent Variable

What is the response variable? Is it quantitative or qualitative? The response variable will be the prediction. It is qualitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative. Category, likes, comments and tags. Likes is quantitative, the others are qualitative.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Summary Statistics

Hide

```
summary(video_df)
```

```
##    video_id          trending_date         title           channel_title     
##  Length:40949       Length:40949       Length:40949       Length:40949      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##    category          publish_time                     tags          
##  Length:40949       Min.   :2006-07-23 08:24:11   Length:40949      
##  Class :character   1st Qu.:2017-12-27 21:00:00   Class :character  
##  Mode  :character   Median :2018-02-21 16:19:27   Mode  :character  
##                     Mean   :2018-02-11 01:00:49                     
##                     3rd Qu.:2018-04-16 17:20:26                     
##                     Max.   :2018-06-14 01:31:53                     
##      views               likes            dislikes       comment_count    
##  Min.   :      549   Min.   :      0   Min.   :      0   Min.   :      0  
##  1st Qu.:   242329   1st Qu.:   5424   1st Qu.:    202   1st Qu.:    614  
##  Median :   681861   Median :  18091   Median :    631   Median :   1856  
##  Mean   :  2360785   Mean   :  74267   Mean   :   3711   Mean   :   8447  
##  3rd Qu.:  1823157   3rd Qu.:  55417   3rd Qu.:   1938   3rd Qu.:   5755  
##  Max.   :225211923   Max.   :5613827   Max.   :1674420   Max.   :1361580  
##  comments_disabled ratings_disabled video_error_or_removed description       
##  Mode :logical     Mode :logical    Mode :logical          Length:40949      
##  FALSE:40316       FALSE:40780      FALSE:40926            Class :character  
##  TRUE :633         TRUE :169        TRUE :23               Mode  :character  
##                                                                              
##                                                                              
##                                                                              
```

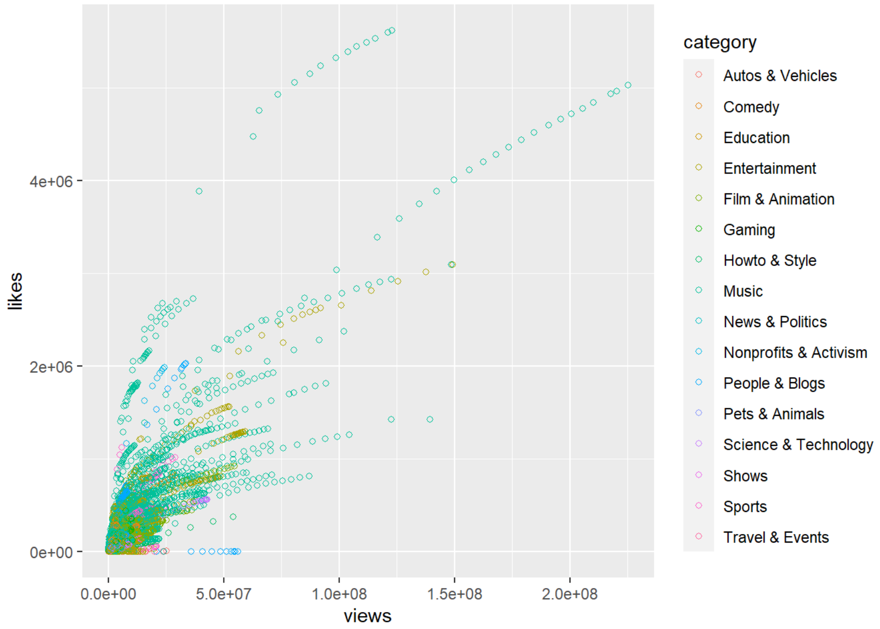
Hide

```
describe(video_df %>% select(views, likes, dislikes))
```

```
##          vars     n      mean         sd median    trimmed       mad min
## views       1 40949 2360784.6 7394113.76 681861 1054836.27 813077.11 549
## likes       2 40949   74266.7  228885.34  18091   32156.33  23496.24   0
## dislikes    3 40949    3711.4   29029.71    631    1137.46    797.64   0
##                max     range  skew kurtosis       se
## views    225211923 225211374 12.24   232.34 36539.66
## likes      5613827   5613827 10.92   177.82  1131.09
## dislikes   1674420   1674420 40.19  1987.08   143.46
```

Hide

```
ggplot(video_df, aes(x=views, y=likes, color = category)) +
    geom_point(shape=1)
```



We see a clear tendency of some categories to gather more views than others.

Hide

```
video_categories <- video_df %>%
  group_by(category) %>%
  summarise(
         views_sum = sum(views),
         likes_sum = sum(likes),
         dislikes_sum = sum(dislikes))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Hide

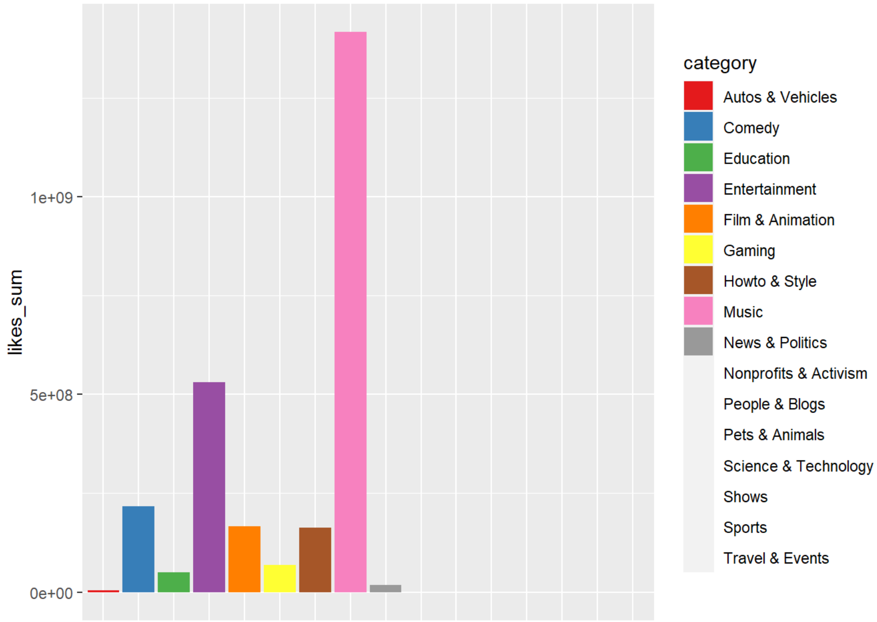
```
knitr::kable(video_categories)
```

category	views_sum	likes_sum	dislikes_sum
Autos & Vehicles	520690717	4245656	243010
Comedy	5117426208	216346746	7230391
Education	1180629990	49257772	1351972
Entertainment	20604388195	530516491	42987663
Film & Animation	7284156721	165997476	6075148
Gaming	2141218625	69038284	9184466
Howto & Style	4078545064	162880075	5473899
Music	40132892190	1416838584	51179008
News & Politics	1473765704	18151033	4180049
Nonprofits & Activism	168941392	14815646	3310381
People & Blogs	4917191726	186615999	10187901
Pets & Animals	764651989	19370702	527379
Science & Technology	3487756816	82532638	4548402
Shows	51501058	1082639	24508
Sports	4404456673	98621211	5133551
Travel & Events	343557084	4836246	340427

Hide

```
ggplot(video_categories, aes(factor(category), likes_sum, fill = category)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1") +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Set1 is 9
## Returning the palette you asked for with that many colors
```



We can see the *Music* category seems to be the one gathering more likes. Further analysis is needed to identify and analyse the tags associated with the different videos and how the presence of these tags might help answer the initial question. …