Project.R

jayshree jagathi

Thu Oct 04 09:13:27 2018

rm(list=ls())  
path<-"F:/DMBI R/Mini Project"  
setwd(path)  
getwd()

## [1] "F:/DMBI R/Mini Project"

df <- read.csv("USvideos.csv")  
str(df)

## 'data.frame': 40949 obs. of 16 variables:  
## $ video\_id : Factor w/ 6351 levels "-\_jlqATo9eo",..: 447 388 780 4406 1824 2487 523 3866 3132 5224 ...  
## $ trending\_date : Factor w/ 205 levels "17.01.12","17.02.12",..: 14 14 14 14 14 14 14 14 14 14 ...  
## $ title : Factor w/ 6455 levels "'Avengers: Infinity War' Cast Tours Los Angeles w/ James Corden",..: 6046 5527 4441 4068 2640 161 4625 256 5352 6305 ...  
## $ channel\_title : Factor w/ 2207 levels "12 News","1MILLION Dance Studio",..: 332 1109 1651 768 1424 890 1681 464 4 2126 ...  
## $ category\_id : int 22 24 23 24 24 28 24 28 1 25 ...  
## $ publish\_time : Factor w/ 6269 levels "2006-07-23T08:24:11.000Z",..: 318 287 271 291 269 323 256 274 297 295 ...  
## $ tags : Factor w/ 6055 levels "#guitar #musiciseverywhere #jammin #meme #funny #deeppurple #pinkfloyd",..: 4673 3071 4300 4442 4521 2487 4808 64 5563 5782 ...  
## $ views : int 748374 2418783 3191434 343168 2095731 119180 2103417 817732 826059 256426 ...  
## $ likes : int 57527 97185 146033 10172 132235 9763 15993 23663 3543 12654 ...  
## $ dislikes : int 2966 6146 5339 666 1989 511 2445 778 119 1363 ...  
## $ comment\_count : int 15954 12703 8181 2146 17518 1434 1970 3432 340 2368 ...  
## $ thumbnail\_link : Factor w/ 6352 levels "https://i.ytimg.com/vi/-\_jlqATo9eo/default.jpg",..: 447 388 780 4407 1824 2487 523 3866 3132 5225 ...  
## $ comments\_disabled : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...  
## $ ratings\_disabled : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...  
## $ video\_error\_or\_removed: Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...  
## $ description : Factor w/ 6902 levels "","'A curious cat helps his owner with home improvements.'\\nWe're releasing a NEW BLACK & WHITE episode every wee"| \_\_truncated\_\_,..: 4844 4286 6380 6122 2630 6262 1611 2724 2973 1872 ...

head(df)

## video\_id trending\_date  
## 1 2kyS6SvSYSE 17.14.11  
## 2 1ZAPwfrtAFY 17.14.11  
## 3 5qpjK5DgCt4 17.14.11  
## 4 puqaWrEC7tY 17.14.11  
## 5 d380meD0W0M 17.14.11  
## 6 gHZ1Qz0KiKM 17.14.11  
## title  
## 1 WE WANT TO TALK ABOUT OUR MARRIAGE  
## 2 The Trump Presidency: Last Week Tonight with John Oliver (HBO)  
## 3 Racist Superman | Rudy Mancuso, King Bach & Lele Pons  
## 4 Nickelback Lyrics: Real or Fake?  
## 5 I Dare You: GOING BALD!?  
## 6 2 Weeks with iPhone X  
## channel\_title category\_id publish\_time  
## 1 CaseyNeistat 22 2017-11-13T17:13:01.000Z  
## 2 LastWeekTonight 24 2017-11-13T07:30:00.000Z  
## 3 Rudy Mancuso 23 2017-11-12T19:05:24.000Z  
## 4 Good Mythical Morning 24 2017-11-13T11:00:04.000Z  
## 5 nigahiga 24 2017-11-12T18:01:41.000Z  
## 6 iJustine 28 2017-11-13T19:07:23.000Z  
## tags  
## 1 SHANtell martin  
## 2 last week tonight trump presidency|last week tonight donald trump|john oliver trump|donald trump  
## 3 racist superman|rudy|mancuso|king|bach|racist|superman|love|rudy mancuso poo bear black white official music video|iphone x by pineapple|lelepons|hannahstocking|rudymancuso|inanna|anwar|sarkis|shots|shotsstudios|alesso|anitta|brazil|Getting My Driver's License | Lele Pons  
## 4 rhett and link|gmm|good mythical morning|rhett and link good mythical morning|good mythical morning rhett and link|mythical morning|Season 12|nickelback lyrics|nickelback lyrics real or fake|nickelback|nickelback songs|nickelback song|rhett link nickelback|gmm nickelback|lyrics (website category)|nickelback (musical group)|rock|music|lyrics|chad kroeger|canada|music (industry)|mythical|gmm challenge|comedy|funny|challenge  
## 5 ryan|higa|higatv|nigahiga|i dare you|idy|rhpc|dares|no truth|comments|comedy|funny|stupid|fail  
## 6 ijustine|week with iPhone X|iphone x|apple|iphone|iphone x review|iphone x unboxing  
## views likes dislikes comment\_count  
## 1 748374 57527 2966 15954  
## 2 2418783 97185 6146 12703  
## 3 3191434 146033 5339 8181  
## 4 343168 10172 666 2146  
## 5 2095731 132235 1989 17518  
## 6 119180 9763 511 1434  
## thumbnail\_link comments\_disabled  
## 1 https://i.ytimg.com/vi/2kyS6SvSYSE/default.jpg False  
## 2 https://i.ytimg.com/vi/1ZAPwfrtAFY/default.jpg False  
## 3 https://i.ytimg.com/vi/5qpjK5DgCt4/default.jpg False  
## 4 https://i.ytimg.com/vi/puqaWrEC7tY/default.jpg False  
## 5 https://i.ytimg.com/vi/d380meD0W0M/default.jpg False  
## 6 https://i.ytimg.com/vi/gHZ1Qz0KiKM/default.jpg False  
## ratings\_disabled video\_error\_or\_removed  
## 1 False False  
## 2 False False  
## 3 False False  
## 4 False False  
## 5 False False  
## 6 False False  
## description  
## 1 SHANTELL'S CHANNEL - https://www.youtube.com/shantellmartin\\nCANDICE - https://www.lovebilly.com\\n\\nfilmed this video in 4k on this -- http://amzn.to/2sTDnRZ\\nwith this lens -- http://amzn.to/2rUJOmD\\nbig drone - http://tinyurl.com/h4ft3oy\\nOTHER GEAR --- http://amzn.to/2o3GLX5\\nSony CAMERA http://amzn.to/2nOBmnv\\nOLD CAMERA; http://amzn.to/2o2cQBT\\nMAIN LENS; http://amzn.to/2od5gBJ\\nBIG SONY CAMERA; http://amzn.to/2nrdJRO\\nBIG Canon CAMERA; http://tinyurl.com/jn4q4vz\\nBENDY TRIPOD THING; http://tinyurl.com/gw3ylz2\\nYOU NEED THIS FOR THE BENDY TRIPOD; http://tinyurl.com/j8mzzua\\nWIDE LENS; http://tinyurl.com/jkfcm8t\\nMORE EXPENSIVE WIDE LENS; http://tinyurl.com/zrdgtou\\nSMALL CAMERA; http://tinyurl.com/hrrzhor\\nMICROPHONE; http://tinyurl.com/zefm4jy\\nOTHER MICROPHONE; http://tinyurl.com/jxgpj86\\nOLD DRONE (cheaper but still great);http://tinyurl.com/zcfmnmd\\n\\nfollow me; on http://instagram.com/caseyneistat\\non https://www.facebook.com/cneistat\\non https://twitter.com/CaseyNeistat\\n\\namazing intro song by https://soundcloud.com/discoteeth\\n\\nad disclosure. THIS IS NOT AN AD. not selling or promoting anything. but samsung did produce the Shantell Video as a 'GALAXY PROJECT' which is an initiative that enables creators like Shantell and me to make projects we might otherwise not have the opportunity to make. hope that's clear. if not ask in the comments and i'll answer any specifics.  
## 2 One year after the presidential election, John Oliver discusses what we've learned so far and enlists our catheter cowboy to teach Donald Trump what he hasn't.\\n\\nConnect with Last Week Tonight online...\\n\\nSubscribe to the Last Week Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/user/LastWeekTonight\\n\\nFind Last Week Tonight on Facebook like your mom would: http://Facebook.com/LastWeekTonight\\n\\nFollow us on Twitter for news about jokes and jokes about news: http://Twitter.com/LastWeekTonight\\n\\nVisit our official site for all that other stuff at once: http://www.hbo.com/lastweektonight  
## 3 WATCH MY PREVIOUS VIDEO â–¶ \\n\\nSUBSCRIBE â–º https://www.youtube.com/channel/UC5jkXpfnBhlDjqh0ir5FsIQ?sub\_confirmation=1\\n\\nTHANKS FOR WATCHING! LIKE & SUBSCRIBE FOR MORE VIDEOS!\\n-----------------------------------------------------------\\nFIND ME ON: \\nInstagram | http://instagram.com/rudymancuso\\nTwitter | http://twitter.com/rudymancuso\\nFacebook | http://facebook.com/rudymancuso\\n\\nCAST: \\nRudy Mancuso | http://youtube.com/c/rudymancuso\\nLele Pons | http://youtube.com/c/lelepons\\nKing Bach | https://youtube.com/user/BachelorsPadTv\\n\\nVideo Effects: \\nCaleb Natale | https://instagram.com/calebnatale\\n\\nPA:\\nPaulina Gregory\\n\\n\\nShots Studios Channels:\\nAlesso | https://youtube.com/c/alesso\\nAnitta | http://youtube.com/c/anitta\\nAnwar Jibawi | http://youtube.com/c/anwar\\nAwkward Puppets | http://youtube.com/c/awkwardpuppets\\nHannah Stocking | http://youtube.com/c/hannahstocking\\nInanna Sarkis | http://youtube.com/c/inanna\\nLele Pons | http://youtube.com/c/lelepons\\nMaejor | http://youtube.com/c/maejor\\nMike Tyson | http://youtube.com/c/miketyson \\nRudy Mancuso | http://youtube.com/c/rudymancuso\\nShots Studios | http://youtube.com/c/shots\\n\\n#Rudy\\n#RudyMancuso  
## 4 Today we find out if Link is a Nickelback amateur or a secret Nickelback devotee. GMM #1218\\nDon't miss an all new Ear Biscuits: https://goo.gl/xeZNQt\\nWatch Part 4: https://youtu.be/MhCdiiB8CQg | Watch Part 2: https://youtu.be/7qiOrNao9fg\\nWatch today's episode from the start: http://bit.ly/GMM1218\\n\\nPick up all of the official GMM merch only at https://mythical.store\\n\\nFollow Rhett & Link: \\nInstagram: https://instagram.com/rhettandlink\\nFacebook: https://facebook.com/rhettandlink\\nTwitter: https://twitter.com/rhettandlink\\nTumblr: https://rhettandlink.tumblr.com\\nSnapchat: @realrhettlink\\nWebsite: https://mythical.co/\\n\\nCheck Out Our Other Mythical Channels:\\nGood Mythical MORE: https://youtube.com/goodmythicalmore\\nRhett & Link: https://youtube.com/rhettandlink\\nThis Is Mythical: https://youtube.com/thisismythical\\nEar Biscuits: https://applepodcasts.com/earbiscuits\\n\\nWant to send us something? https://mythical.co/contact\\nHave you made a Wheel of Mythicality intro video? Submit it here: https://bit.ly/GMMWheelIntro\\n\\nIntro Animation by Digital Twigs: https://www.digitaltwigs.com\\nIntro & Outro Music by Jeff Zeigler & Sarah Schimeneck https://www.jeffzeigler.com\\nWheel of Mythicality theme: https://www.royaltyfreemusiclibrary.com/\\nAll Supplemental Music fromOpus 1 Music: https://opus1.sourceaudio.com/\\nWe use â\200\230The Mouseâ\200\231 by Blue Microphones https://www.bluemic.com/mouse/  
## 5 I know it's been a while since we did this show, but we're back with what might be the best episode yet!\\nLeave your dares in the comment section! \\n\\nOrder my book how to write good \\nhttp://higatv.com/ryan-higas-how-to-write-good-pre-order-links/\\n\\nJust Launched New Official Store\\nhttps://www.gianthugs.com/collections/ryan\\n\\nHigaTV Channel\\nhttp://www.youtube.com/higatv\\n\\nTwitter\\nhttp://www.twitter.com/therealryanhiga\\n\\nFacebook\\nhttp://www.facebook.com/higatv\\n\\nWebsite\\nhttp://www.higatv.com\\n\\nInstagram\\nhttp://www.instagram.com/notryanhiga\\n\\nSend us mail or whatever you want here!\\nPO Box 232355\\nLas Vegas, NV 89105  
## 6 Using the iPhone for the past two weeks -- here's my thoughts!\\nAll my iPhone X Videos: https://www.youtube.com/watch?v=vqztGUwhRlQ&list=PLoYRB6C09WUDbCndtEDELX-Fpk\_pgATvF\\nâ–º SUBSCRIBE FOR MORE VIDEOS: http://www.youtube.com/subscription\_center?add\_user=ijustine\\nâ–º Get my BOOK! http://ijustinebook.com\\nâ–º Get my iOS 10 STICKERS! http://ijustinestickers.com\\n\\nâ–¼ SOCIAL\\nhttp://instagram.com/ijustine\\nhttp://facebook.com/ijustine\\nhttp://twitter.com/ijustine\\nSnapchat: iJustine\\n\\nâ\200”â\200”â\200”â\200”â\200”â\200”â\200”â\200”â\200”â\200”â\200”â\200”\\n\\nâ–¼ STUFF I USE TO MAKE VIDEOS\\n\\nSony rx100 V - http://amzn.to/2jesbxA\\nG7X ii - http://amzn.to/2f6n2Bs\\nCanon 80D - http://amzn.to/2eRKhQo\\nSony A7s ii - http://amzn.to/2ebLR16\\nSony a6500 - http://amzn.to/2okeG2a\\nGoPro Hero 5 - http://amzn.to/2e1KyhM\\nGoPro Hero 5 Session - http://amzn.to/2oksMQT\\nEpidemic Sound - https://goo.gl/Pe7GTL \\n\\nFavorite lenses: \\nCanon EF 24-70mm - http://amzn.to/2dT7mFr\\nCanon EF-S 10-18mm - http://amzn.to/2dT62SU\\nSony 16-35mm (full frame) - http://amzn.to/2ftPaTf\\nSony Distagon 35mm (full frame) - http://amzn.to/2oB0XQj\\nSony 10-18mm wide angle - http://amzn.to/2e1Myqz\\n\\nRode Small on camera microphone - http://amzn.to/2fkiVGJ\\nRode Larger (battery required) microphone - http://amzn.to/2ftNkl8\\nSony XLR adapter Microphone - http://amzn.to/2kCcIDH\\nSmall Sony Microphone - http://amzn.to/2oX7Eih\\n\\nFavorite SD Card - http://amzn.to/2oWRGoD\\n\\nDJI Mavic Pro - http://amzn.to/2f6nL5E\\nPhantom 4 Pro - http://amzn.to/2pbDrN1\\nPhantom 4 Pro Plus - http://amzn.to/2oX63Jz

library(plyr)  
  
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::arrange() masks plyr::arrange()  
## x purrr::compact() masks plyr::compact()  
## x dplyr::count() masks plyr::count()  
## x dplyr::failwith() masks plyr::failwith()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::id() masks plyr::id()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::mutate() masks plyr::mutate()  
## x dplyr::rename() masks plyr::rename()  
## x dplyr::summarise() masks plyr::summarise()  
## x dplyr::summarize() masks plyr::summarize()

library(wordcloud)

## Loading required package: RColorBrewer

library(tm)

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

library(SnowballC)  
  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:plyr':  
##   
## here

## The following object is masked from 'package:base':  
##   
## date

library(ggcorrplot)  
  
library(DMwR)

## Loading required package: lattice

## Loading required package: grid

##   
## Attaching package: 'DMwR'

## The following object is masked from 'package:plyr':  
##   
## join

library(caret)

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(rpart)  
library(rpart.plot)  
  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(ipred)  
  
library(caretEnsemble)

##   
## Attaching package: 'caretEnsemble'

## The following object is masked from 'package:ggplot2':  
##   
## autoplot

# removing the thumbnail link column  
df <- df[,-12]  
  
# setting category\_id as a factor variable  
df$category\_id <- as.factor(df$category\_id)  
  
# setting all 0 values for likes, dislikes and comments as NA  
df[, 9:11][df[, 9:11] == 0] <- NA  
  
# counting the missing values  
sapply(df, function(x) sum(is.na(x)))

## video\_id trending\_date title   
## 0 0 0   
## channel\_title category\_id publish\_time   
## 0 0 0   
## tags views likes   
## 0 0 172   
## dislikes comment\_count comments\_disabled   
## 383 760 0   
## ratings\_disabled video\_error\_or\_removed description   
## 0 0 0

# removing the records where the likes and dislikes both are NA  
df <- df %>%   
 filter(!is.na(likes) & !is.na(dislikes))  
  
# imputing the missing comment\_count values using knn  
knnOut <- round(knnImputation(df[,8:11], k = 10))  
  
# inserting the imputed values into the original dataframe  
df <- cbind(df[,1:10], knnOut[,4], df[,12:15])  
  
# renaming the column  
colnames(df)[11] <- 'comment\_count'  
  
# counting the missing values  
sapply(df, function(x) sum(is.na(x)))

## video\_id trending\_date title   
## 0 0 0   
## channel\_title category\_id publish\_time   
## 0 0 0   
## tags views likes   
## 0 0 0   
## dislikes comment\_count comments\_disabled   
## 0 0 0   
## ratings\_disabled video\_error\_or\_removed description   
## 0 0 0

# a new derived variable that gives the ratio of likes and dislikes for a video   
df$like\_percentage <- df$likes/(df$dislikes+df$likes)  
  
# converting the trending date into date format  
df$trending\_date <- ydm(df$trending\_date)  
  
# getting the publish date for the trending video  
df$publish\_date <- ymd(substr(df$publish\_time, start = 1, stop = 10))  
  
# calculating the difference between the video being published and became trending  
df$diff\_days <- df$trending\_date-df$publish\_date  
  
# converting diff\_days into a numerical feature  
df$diff\_days <- as.numeric(df$diff\_days)  
  
# function to narmalize the numerical features  
z\_normalize <- function(x) {  
 return ((x - mean(x)) / sd(x))  
}  
  
# normalizing the numerical features  
youtube\_norm <- cbind(df[,1:7], lapply(df[,c(8:11, 18)], z\_normalize),   
 df[,12:17])  
  
# summary of numerical features after normalizing  
summary(youtube\_norm[,8:12])

## views likes dislikes   
## Min. :-0.32008 Min. :-0.32616 Min. :-0.12843   
## 1st Qu.:-0.28661 1st Qu.:-0.30151 1st Qu.:-0.12123   
## Median :-0.22690 Median :-0.24583 Median :-0.10638   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.07158 3rd Qu.:-0.08221 3rd Qu.:-0.06111   
## Max. :30.15649 Max. :24.09727 Max. :57.28373   
## comment\_count diff\_days   
## Min. :-0.22814 Min. :-0.11215   
## 1st Qu.:-0.21045 1st Qu.:-0.08978   
## Median :-0.17640 Median :-0.07486   
## Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.07026 3rd Qu.:-0.04504   
## Max. :35.98892 Max. :31.31692

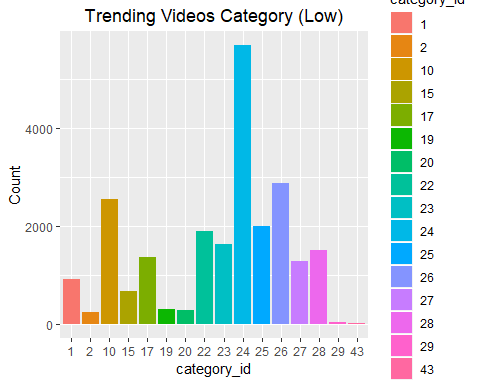
# clearly, the range is too large for one dataset, as we can see that the 3rd quartile  
# for each of the numerical features is less than the mean. and, then there is a small  
# porition of videos with extremely high number of views. so, moving forward, the data will  
# be divided into two subsets, one with views less than one million and the other with  
# views greater than one million and less than ten million  
# videos with views more than 10 million will be removed  
  
# videos with less than one million views  
low\_youtube <- df %>%   
 filter(views < 1000000)  
  
# normalizing the numerical features  
low\_youtube\_norm <- cbind(low\_youtube[,1:7], lapply(low\_youtube[,c(8:11, 18)], z\_normalize),   
 low\_youtube[,12:17])  
  
# removing the rows with sd greater than 3  
low\_youtube\_norm <- low\_youtube\_norm %>%   
 filter(likes < 3 & dislikes < 3 & comment\_count < 3 & diff\_days < 3)  
  
# summary of numerical features after normalizing  
summary(low\_youtube\_norm[,8:12])

## views likes dislikes   
## Min. :-1.32374 Min. :-0.84773 Min. :-0.307263   
## 1st Qu.:-0.86726 1st Qu.:-0.67843 1st Qu.:-0.250682   
## Median :-0.24397 Median :-0.32578 Median :-0.166596   
## Mean :-0.03528 Mean :-0.09011 Mean :-0.058296   
## 3rd Qu.: 0.64981 3rd Qu.: 0.25505 3rd Qu.:-0.007068   
## Max. : 2.28626 Max. : 2.99988 Max. : 2.956642   
## comment\_count diff\_days   
## Min. :-0.57768 Min. :-0.11289   
## 1st Qu.:-0.45130 1st Qu.:-0.09556   
## Median :-0.26675 Median :-0.08402   
## Mean :-0.09964 Mean :-0.07860   
## 3rd Qu.: 0.05567 3rd Qu.:-0.07247   
## Max. : 2.99372 Max. : 2.99943

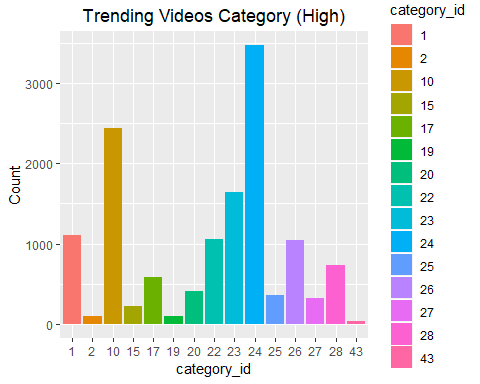
# videos with views between one and ten million  
high\_youtube <- df %>%   
 filter(views > 1000000 & views < 10000000)  
  
# normalizing the numerical features  
high\_youtube\_norm <- cbind(high\_youtube[,1:7], lapply(high\_youtube[,c(8:11, 18)], z\_normalize),   
 high\_youtube[,12:17])  
  
# removing the rows with sd greater than 3  
high\_youtube\_norm <- high\_youtube\_norm %>%   
 filter(likes < 3 & dislikes < 3 & comment\_count < 3 & views < 3 & diff\_days < 3)  
  
# summary of numerical features after normalizing  
summary(high\_youtube\_norm[,8:12])

## views likes dislikes   
## Min. :-0.9269 Min. :-0.8576 Min. :-0.39882   
## 1st Qu.:-0.7364 1st Qu.:-0.5785 1st Qu.:-0.29936   
## Median :-0.4098 Median :-0.3405 Median :-0.21971   
## Mean :-0.1119 Mean :-0.1136 Mean :-0.08599   
## 3rd Qu.: 0.2396 3rd Qu.: 0.1574 3rd Qu.:-0.03350   
## Max. : 2.9983 Max. : 2.9828 Max. : 2.98864   
## comment\_count diff\_days   
## Min. :-0.603905 Min. :-1.48185   
## 1st Qu.:-0.446009 1st Qu.:-0.73832   
## Median :-0.293910 Median :-0.18067   
## Mean :-0.125627 Mean :-0.04432   
## 3rd Qu.: 0.002303 3rd Qu.: 0.56286   
## Max. : 2.943379 Max. : 2.97934

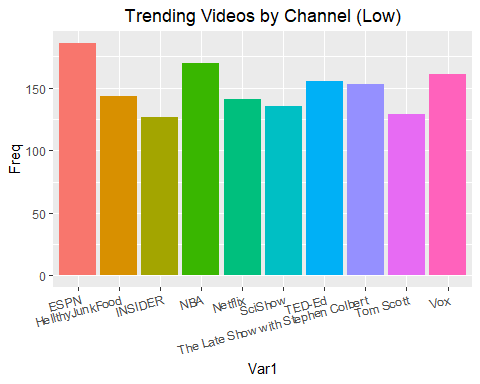
#EDA  
# bar plot to see the most popular cateogry for trending videos  
ggplot(low\_youtube\_norm) +  
 geom\_bar(aes(category\_id, fill = category\_id)) +  
 labs(y = 'Count', title = 'Trending Videos Category (Low)') +  
 theme(plot.title = element\_text(hjust = 0.5))



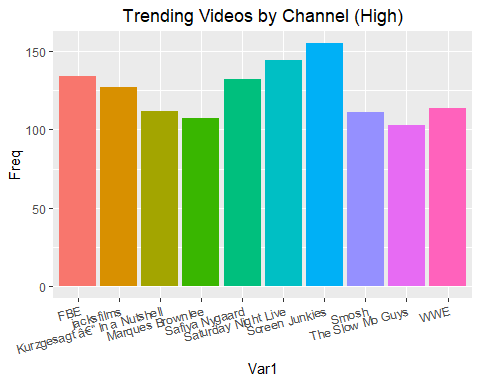
# for videos with views < one million category 24 is the most popular,  
# followed by 10 and 26  
ggplot(high\_youtube\_norm) +  
 geom\_bar(aes(category\_id, fill = category\_id)) +  
 labs(y = 'Count', title = 'Trending Videos Category (High)') +  
 theme(plot.title = element\_text(hjust = 0.5))



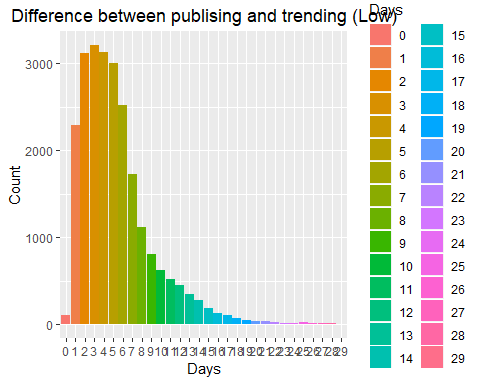
# for videos with views > one million, again category 24 is the most  
# popular, with category 10 being the second most frequent  
# creating a frequency table to determine the number of trending videos by channel  
low\_channels <- as.data.frame(table(low\_youtube\_norm$channel\_title))  
low\_channel\_freq <- low\_channels %>%   
 filter(rank(desc(Freq)) <= 10)  
  
# bar plot to visualize the top channels by number of trending videos  
ggplot(low\_channel\_freq) +  
 geom\_bar(aes(Var1, Freq, fill = Var1), stat = 'identity') +  
 labs(title = 'Trending Videos by Channel (Low)') +  
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle = 15, hjust = 1),  
 legend.position="none")



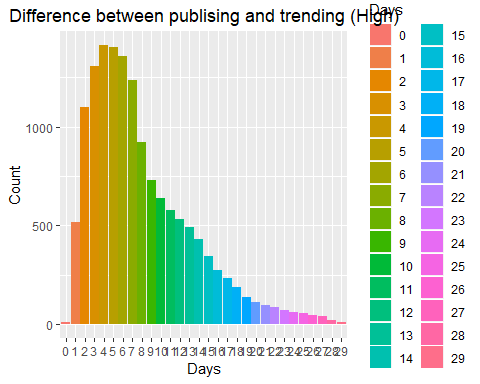
# ESPN has the highest number of trending videos with views less than one million  
# creating a frequency table to determine the number of trending videos by channel  
high\_channels <- as.data.frame(table(high\_youtube\_norm$channel\_title))  
high\_channel\_freq <- high\_channels %>%   
 filter(rank(desc(Freq)) <= 10)  
  
# bar plot to visualize the top channels by number of trending videos  
ggplot(high\_channel\_freq) +  
 geom\_bar(aes(Var1, Freq, fill = Var1), stat = 'identity') +  
 labs(title = 'Trending Videos by Channel (High)') +  
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle = 15, hjust = 1),  
 legend.position="none")



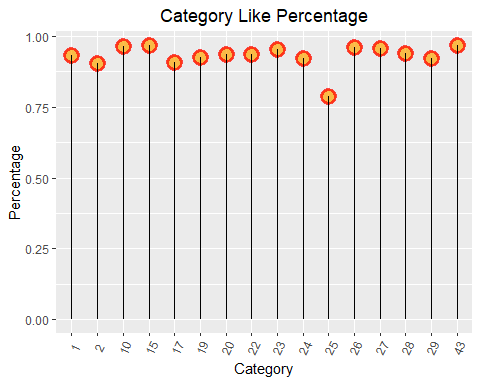
# Screen Junkies has the highest number of trending videos with views more than  
# one million  
# difference between publishing and trending days  
ggplot(subset(low\_youtube, diff\_days<30)) +  
 geom\_bar(aes(as.factor(diff\_days), fill = as.factor(diff\_days))) +  
 labs(x = 'Days', y = 'Count', title = 'Difference between publising and trending (Low)') +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 guides(fill=guide\_legend(title="Days"))



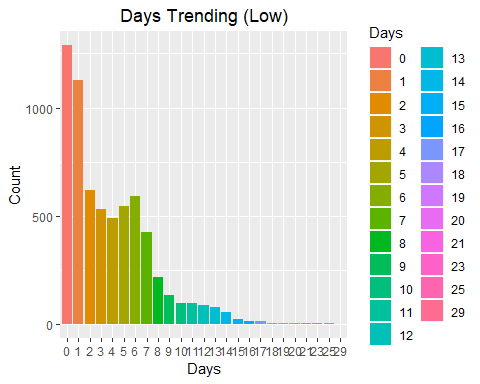
# low view count video will most likely start trending within 10 days   
# of being published  
ggplot(subset(high\_youtube, diff\_days<30)) +  
 geom\_bar(aes(as.factor(diff\_days), fill = as.factor(diff\_days))) +  
 labs(x = 'Days', y = 'Count', title = 'Difference between publising and trending (High)') +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 guides(fill=guide\_legend(title="Days"))



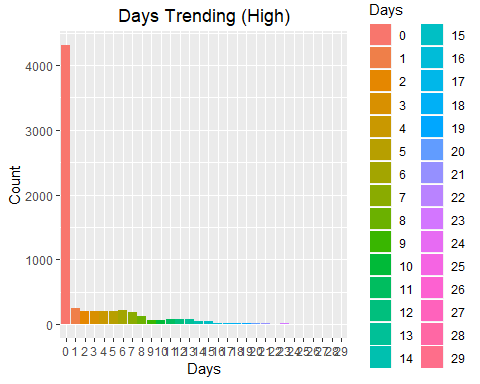
# high view count video will also most likely start trending within   
# 10 days of being published  
# calculating the category wise like percentage of videos  
cat\_like\_mean <- ddply(df, .(category\_id), summarize,   
 mean = mean(like\_percentage, na.rm = TRUE))  
  
# plotting the category wise like percentage  
ggplot(cat\_like\_mean, aes(x = category\_id, y = mean)) +   
 geom\_point(size = 4, color = "red", fill = alpha("orange", 0.3),   
 alpha = 0.7, shape = 21, stroke = 2) +   
 geom\_segment(aes(x = category\_id, xend = category\_id, y = 0, yend = mean)) +   
 labs(title="Category Like Percentage", x = 'Category', y = 'Percentage') +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6),   
 plot.title = element\_text(hjust = 0.5))



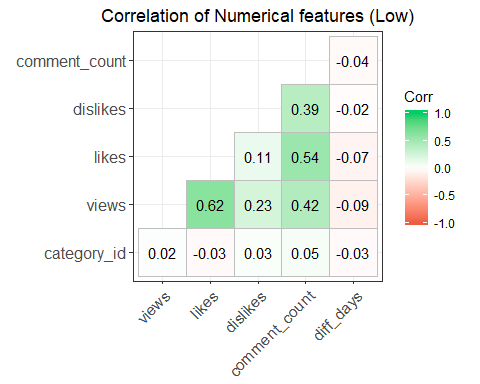
# although the category 24 has the highest number of trending videos, the   
# ratio for number of likes and dislikes is the lowest  
# graph to see for how long does a video remain trending for low view count  
low\_days\_trending <- as.data.frame(table(low\_youtube$title))  
  
ggplot(low\_days\_trending, aes(as.factor(Freq), fill = as.factor(Freq))) +  
 geom\_bar() +  
 labs(x = 'Days', y = 'Count', title = 'Days Trending (Low)') +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 guides(fill=guide\_legend(title="Days"))



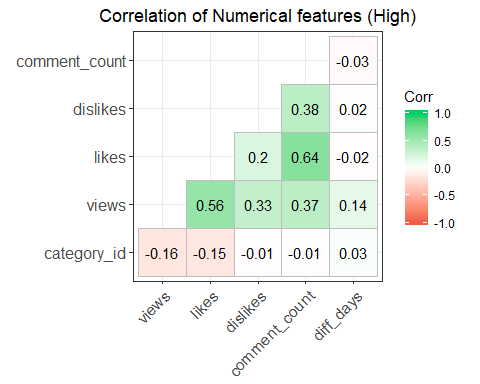
# a video with lower number of views can remain trending for more than a week  
# graph to see for how long does a video remain trending for high view count  
high\_days\_trending <- as.data.frame(table(high\_youtube$title))  
  
ggplot(high\_days\_trending, aes(as.factor(Freq), fill = as.factor(Freq))) +  
 geom\_bar() +  
 labs(x = 'Days', y = 'Count', title = 'Days Trending (High)') +  
 theme(plot.title = element\_text(hjust = 0.5)) +   
 guides(fill=guide\_legend(title="Days"))



# a video with higher view count is not likely to remain trending for more than  
# a single day  
# getting the numerical features of the data frame for low video count  
low\_num\_youtube <- low\_youtube[,c("category\_id","views","likes","dislikes",  
 "comment\_count", 'diff\_days')]  
  
# setting categorical\_id as numerical feature  
low\_num\_youtube$category\_id <- as.numeric(low\_num\_youtube$category\_id)  
  
# calculating the correlations among numerical features  
low\_cor\_num\_youtube <- cor(low\_num\_youtube, use = 'pairwise.complete.obs')  
  
# plot for correlations among numerical features  
ggcorrplot(low\_cor\_num\_youtube, hc.order = FALSE, type = "lower", lab = TRUE,   
 lab\_size = 4, colors = c("tomato2", "white", "springgreen3"),   
 title = "Correlation of Numerical features (Low)",   
 ggtheme = theme\_bw) +  
 theme(plot.title = element\_text(hjust = 0.5))



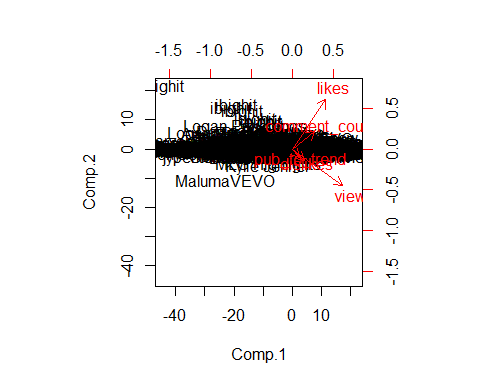
# we can see that number of views, likes, dislikes and the comment count   
# for a video are highly correlated  
# repeating the procedure for videos with high view count  
high\_num\_youtube <- high\_youtube[,c("category\_id","views","likes","dislikes",  
 "comment\_count", 'diff\_days')]  
  
high\_num\_youtube$category\_id <- as.numeric(high\_num\_youtube$category\_id)  
  
high\_cor\_num\_youtube <- cor(high\_num\_youtube, use = 'pairwise.complete.obs')  
  
ggcorrplot(high\_cor\_num\_youtube, hc.order = FALSE, type = "lower", lab = TRUE,   
 lab\_size = 4, colors = c("tomato2", "white", "springgreen3"),   
 title = "Correlation of Numerical features (High)",   
 ggtheme = theme\_bw) +  
 theme(plot.title = element\_text(hjust = 0.5))



# again number of views, likes, dislikes and the comment count for a   
# video are highly correlated  
  
###############################################################################  
  
  
  
utube\_us <- read.csv("USvideos.csv",   
 encoding = "UTF-8", stringsAsFactors=FALSE,  
 na.strings=c("", "NA"))  
  
utube\_us$category\_id <- factor(utube\_us$category\_id)  
utube\_us$video\_id <- factor(utube\_us$video\_id)  
utube\_us$channel\_title <- factor(utube\_us$channel\_title)  
utube\_us$comments\_disabled <- factor(utube\_us$comments\_disabled)  
utube\_us$ratings\_disabled <- factor(utube\_us$ratings\_disabled)  
utube\_us$video\_error\_or\_removed <- factor(utube\_us$video\_error\_or\_removed)  
utube\_us$trending\_date <- as.Date(utube\_us$trending\_date, format = '%y.%d.%m')  
utube\_us$publish\_time <- as.Date(utube\_us$publish\_time, format = '%Y-%m-%d')  
utube\_us$pub\_to\_trend <- as.numeric(utube\_us$trending\_date - utube\_us$publish\_time)  
# description column has missing values  
missing\_per\_col <- sapply(utube\_us, function(x) sum(is.na(x)))  
(total\_missing <- sum(missing\_per\_col))

## [1] 570

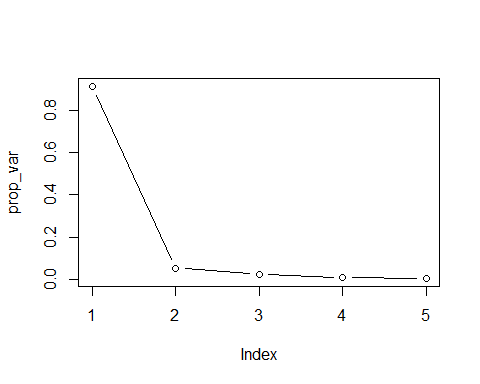
# clean discription column  
# exclude emojis  
utube\_us\_nodup <- utube\_us[!duplicated(utube\_us$video\_id), ]  
  
# PCA  
# channels associated with each other  
  
library(pcaPP)  
nodup\_numeric <-select\_if(utube\_us\_nodup, is.numeric)  
pr\_out <- PCAproj(nodup\_numeric, scale = sd, k = 5)  
rownames(pr\_out$scores) <- utube\_us\_nodup$channel\_title  
biplot(pr\_out, scale = 0)



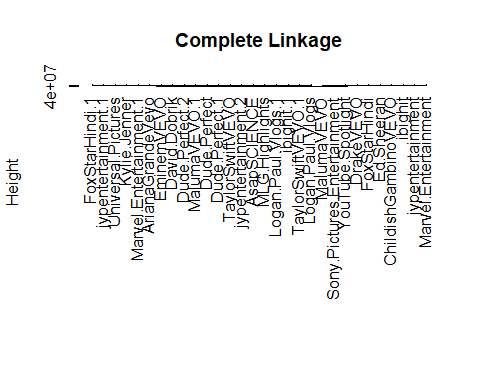
pr\_out$loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## views 0.766 -0.558 -0.277 -0.132   
## likes 0.506 0.754 -0.288 0.292   
## dislikes 0.183 -0.173 0.439 0.320 0.801  
## comment\_count 0.337 0.270 0.705 -0.528 -0.194  
## pub\_to\_trend 0.102 -0.133 0.389 0.718 -0.552  
##   
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## SS loadings 1.0 1.0 1.0 1.0 1.0  
## Proportion Var 0.2 0.2 0.2 0.2 0.2  
## Cumulative Var 0.2 0.4 0.6 0.8 1.0

prop\_var <- (pr\_out$sdev ^ 2) / (sum(pr\_out$sdev ^ 2))  
plot(prop\_var, type='b')



# hierarchical clustering  
# so that the channel names show up in the plot instead of numbers  
top\_views <- nodup\_numeric %>%   
 mutate(channel\_title = utube\_us\_nodup$channel\_title) %>%  
 arrange(desc(views))  
rownames(top\_views) <- make.names(top\_views$channel\_title, unique = TRUE)  
top\_views <- select(top\_views, -channel\_title)  
  
hc\_complete <- hclust(dist(top\_views[1:30, ]), method = "complete")  
plot(hc\_complete, main = "Complete Linkage", xlab = "", sub = "")



which(cutree(hc\_complete, 6) == 5)

## Kylie.Jenner Marvel.Entertainment.1 FoxStarHindi.1   
## 11 12 13   
## jypentertainment.1 Universal.Pictures   
## 14 15