Xinyi & Daiyan 632 Project

```
library(tidyverse)
library(stringr)
```

Inspect the data

Load the data:

```
df_raw <- read.csv("data.csv")
head(df_raw)
```

```
##            Id              Channel       Subscribers
## 1 FozCkl1xj-w            JRE Clips 6.28M subscribers
## 2 RN8yoi-e2yc     Mythical Kitchen  1.9M subscribers
## 3 IugcIAAZJ2M             Munchies 4.59M subscribers
## 4 JiEO6F8i0eU Parks and Recreation  282K subscribers
## 5 1T4XMNN4bNM               Vsauce 17.4M subscribers
## 6 0ZWGeidvrJw           Doctor Who 1.59M subscribers
##                                                                    Title CC
## 1                        Former CIA Agent Breaks Down Jeffrey Epstein Case  0
## 2 $420 Pizza Hut Stuffed Crust Pizza | Fancy Fast Food | Mythical Kitchen  1
## 3                    The Iconic $1 Pizza Slice of NYC | Street Food Icons  0
## 4                   Ron Swanson: The Papa of Pawnee | Parks and Recreation  0
## 5                              What's The Most Dangerous Place on Earth?  1
## 6     The Doctor Defeats the Abzorbaloff | Love and Monsters | Doctor Who  1
##                                          URL    Released      Views
## 1 https://www.youtube.com/watch?v=FozCkl1xj-w 2 years ago 7.9M views
## 2 https://www.youtube.com/watch?v=RN8yoi-e2yc             2.7M views
## 3 https://www.youtube.com/watch?v=IugcIAAZJ2M 2 years ago  11M views
## 4 https://www.youtube.com/watch?v=JiEO6F8i0eU 3 years ago 2.3M views
## 5 https://www.youtube.com/watch?v=1T4XMNN4bNM 9 years ago  21M views
## 6 https://www.youtube.com/watch?v=0ZWGeidvrJw 7 years ago 8.5M views
##               Category
## 1                  Blog
## 2                  Food
## 3                  Food
## 4 Entertainment,Comedy
## 5               Science
## 6         Entertainment
##
## 1
## 2 - Oh, that's dirty.\n- Wow! - Whoa.\n- You're a dirty girl. (upbeat music) - Hey man. - What'd you
## 3
## 4
## 5
```

```
## 6
##   Length
## 1  13:32
## 2  24:26
## 3   7:51
## 4  10:06
## 5   9:29
## 6   4:20
```

We notice several problems, like:

1. Views: The Views variable is in string format and the units are different, like "10K views", "10M views". We prefer it to be in number and in the same unit in order to conduct statistical analysis.
2. Subscribers: The same problems as Views. The Subscribers variable is like "10K subscribers", "10M subscribers".
3. Length: The video length is in string format, like "12:00", "1:12:00". We need it to be in number and in the same unit.
4. Released: The Released variable is in string format, like "2 years age", "10 month ago". We need it to be in number and in the same unit.

Therefore, we need to do data cleaning first.

Data Cleaning

```
# Unify units and convert string to number, like: 10K views -> 10, 10M views -> 10000
cleanViews <- function(str) {
  str <- str_remove(str, " views")
  last <- str_sub(str, -1)
  views <- str %>% str_remove(last) %>% as.numeric()
  if (last == "M") return(1000*views)
  else return(views)
}

cleanViewsCol <-  function(col) {
  ret <- c()
  for(r in col) {
    ret <- append(ret, cleanViews(r))
  }

  return(ret)
}

# Unify units and convert string to number, like: 10K subscribers -> 10, 10M subscribers -> 10000
cleanSubscribers <- function(str) {
  str <- str_remove(str, " subscribers")
  last <- str_sub(str, -1)
  views <- str %>% str_remove(last) %>% as.numeric()
  if (last == "M") return(1000*views)
  else return(views)
```

```
}

cleanSubscribersCol <-  function(col) {
  ret <- c()
  for(r in col) {
    ret <- append(ret, cleanSubscribers(r))
  }
  return(ret)
}

# Convert time in string format to number of minutes, like: 12:00 -> 12, 1:12:00 -> 72
cleanLength <- function(str) {
  list <- str_split(str, ":")
  len <- length(list[[1]])
  if (len == 3) {
     h <- as.numeric(list[[1]][1])
     m <- as.numeric(list[[1]][2])
     return((m + 1) + 60*m)
  } else {
     m <- as.numeric(list[[1]][1])
     return(m+1)
  }
}

cleanLengthCol <-  function(col) {
  ret <- c()
  for(r in col) {
    ret <- append(ret, cleanLength(r))
  }
  return(ret)
}

# Convert time to number of months ago, like: 1 years ago -> 12, 10 months ago to 10
cleanReleased <- function(str) {
  str <- str_remove(str, "Streamed ")
  list <- str_split(str, " ")
  if (list[[1]][2] == "years") return(as.numeric(list[[1]][1])*12)
  else return(as.numeric(list[[1]][1]))
}

cleanReleasedCol <- function(col) {
  ret <- c()
  for(r in col) {
    ret <- append(ret, cleanReleased(r))
  }
  return(ret)
}

# Remove NAs
df <- df_raw %>%
  filter(
    !is.na(Released) & Released != ""
  )
```

```
# Clean the data
df$Released = df$Released %>% cleanReleasedCol()
df$Subscribers = df$Subscribers %>% cleanSubscribersCol()
df$Views = df$Views %>% cleanViewsCol()
df$Length = df$Length %>% cleanLengthCol()

head(df)

# Save for future use
write.csv(df, "cleaned_data.csv")
```

Check the cleaned data:

```
df <- read.csv("cleaned_data.csv")
head(df)
```

```
##   X          Id              Channel Subscribers
## 1 1 FozCkl1xj-w           JRE Clips        6280
## 2 2 IugcIAAZJ2M             Munchies        4590
## 3 3 JiEO6F8i0eU Parks and Recreation         282
## 4 4 1T4XMNN4bNM               Vsauce       17400
## 5 5 0ZWGeidvrJw           Doctor Who        1590
## 6 6 YiEj9mrqTN0                 A&E        7930
##                                                                    Title CC
## 1                   Former CIA Agent Breaks Down Jeffrey Epstein Case  0
## 2                The Iconic $1 Pizza Slice of NYC | Street Food Icons  0
## 3              Ron Swanson: The Papa of Pawnee | Parks and Recreation  0
## 4                           What's The Most Dangerous Place on Earth?  1
## 5 The Doctor Defeats the Abzorbaloff | Love and Monsters | Doctor Who  1
## 6           Live PD: Most Viewed Moments from Walton County, FL | A&E  1
##                                          URL Released Views
## 1 https://www.youtube.com/watch?v=FozCkl1xj-w       24  7900
## 2 https://www.youtube.com/watch?v=IugcIAAZJ2M       24 11000
## 3 https://www.youtube.com/watch?v=JiEO6F8i0eU       36  2300
## 4 https://www.youtube.com/watch?v=1T4XMNN4bNM      108 21000
## 5 https://www.youtube.com/watch?v=0ZWGeidvrJw       84  8500
## 6 https://www.youtube.com/watch?v=YiEj9mrqTN0       24 14000
##               Category
## 1                 Blog
## 2                 Food
## 3 Entertainment,Comedy
## 4              Science
## 5        Entertainment
## 6                 News
##
## 1
r lunch they k##w2they gotta just keep making them as the people are coming so they know when they're gonna get hit a
es don't worry##t3s gonna be fine ow if you don't take care of the problem now it's only gonna get worse come on I'll
 highest murde##r4te you'll have\nto go to Juarez, Mexico, where out of every 1 million inhabitants, each year 1,477\
tle man. Come o##.5There's everlasting peace. Come on. join us. Dissolve into me. Someone wants a word with you. You u
## 6 [music playing] We'll be on Laverne Street. It's a red Dodge pickup. This whole not stopping\nthin
##   Length
```

```
## 1     14
## 2      8
## 3     11
## 4     10
## 5      5
## 6     21
```

After cleaning, Views, Subscribers, Released and Length are numbers, while Views, Subscribers are in K and Released and Length are in minute.

Sentiment Analysis

TODO by Xinyi

Linear Regression

TODO by Xinyi

Random Forest

TODO by Daiyan

Conclusion

TODO