Predictive Model for YouTrend

Daivaksh Patel

```
# load necessary packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(gridExtra)
```

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

```
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Gathering and cleaning/validating dataset

```
# Load the 'grades.csv' data set
bra = read.csv("BR_youtube_trending_data.csv")
usa = read.csv("US_youtube_trending_data (3).csv")
can = read.csv("CA_youtube_trending_data.csv")
ind = read.csv("IN_youtube_trending_data.csv")
jpn = read.csv("JP_youtube_trending_data.csv")
```

```
# Looking at the structure of the data
dim(bra)
```

```
## [1] 10000    18
```

```r
dim(usa)
```

```
## [1] 10000    18
```

```r
dim(can)
```

```
## [1] 10000    18
```

```r
dim(ind)
```

```
## [1] 10000    18
```

```r
dim(jpn)
```

```
## [1] 10000    18
```

```r
colnames(bra)
```

```
##  [1] "video_id"          "title"             "publishedAt"
##  [4] "channelId"         "channelTitle"      "categoryId"
##  [7] "trending_date"     "tags"              "view_count"
## [10] "likes"             "dislikes"          "comment_count"
## [13] "thumbnail_link"    "comments_disabled" "ratings_disabled"
## [16] "description"       "countryName"       "countryCode"
```

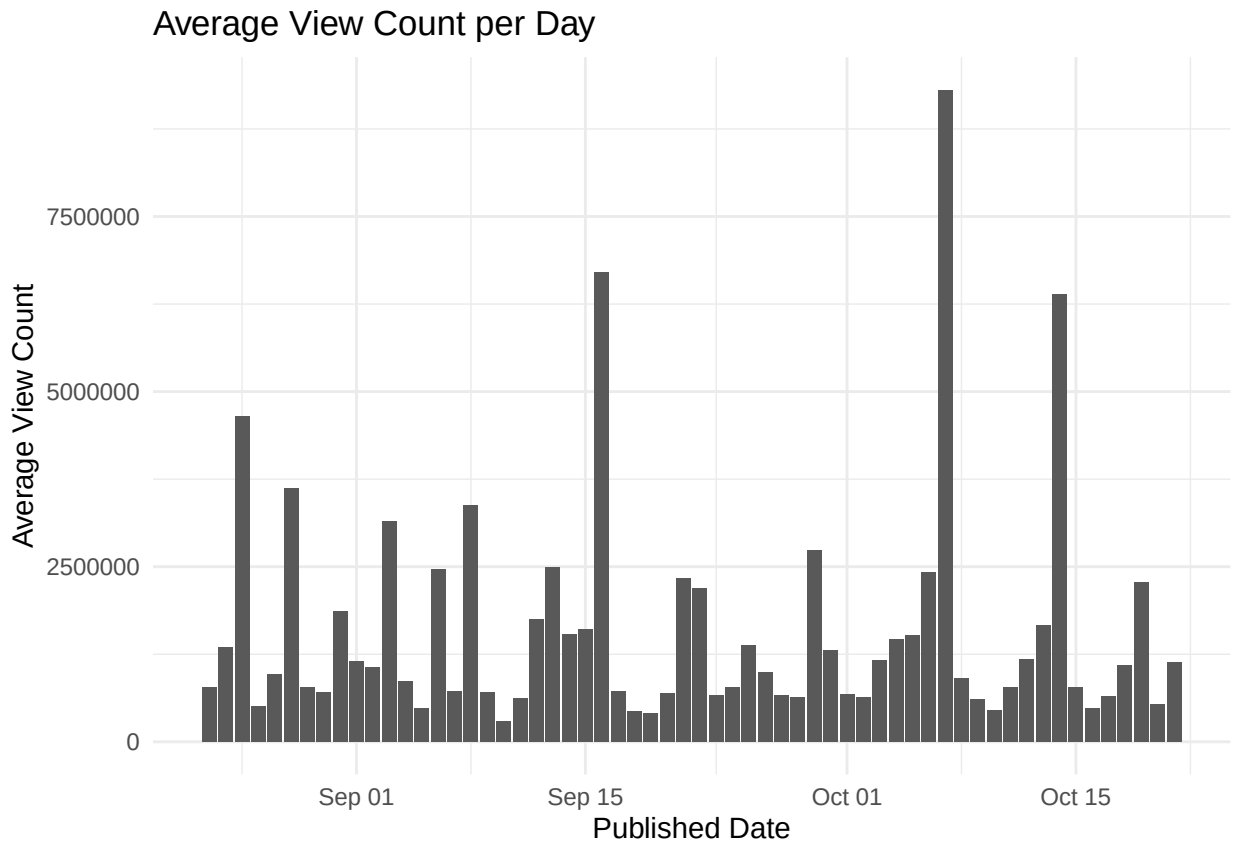
```r
str(bra)
```

```
## 'data.frame':    10000 obs. of  18 variables:
##  $ video_id         : chr  "h09p0IGiKaE" "Mq3QAowjuTk" "wI8x14J_kcw" "S---FS9Dnto" ...
##  $ title            : chr  "Gusttavo Lima - Canudinho Part. Ana Castela | DVD Paraíso Particular" "J
##  $ publishedAt      : chr  "2023-09-01T13:59:57Z" "2023-09-01T20:58:14Z" "2023-09-01T20:54:58Z" "202
##  $ channelId        : chr  "UCXooz9whNJZBRTHi9AqdjPw" "UC_oToDrJ6uca7d1dFVBmLtg" "UC_oToDrJ6uca7d1dF
##  $ channelTitle     : chr  "Gusttavo Lima Oficial" "Canal GOAT" "Canal GOAT" "BETO GAMER" ...
##  $ categoryId       : int  10 17 17 20 17 10 24 10 10 10 ...
##  $ trending_date    : chr  "2023-09-02T00:00:00Z" "2023-09-02T00:00:00Z" "2023-09-02T00:00:00Z" "202
##  $ tags             : chr  "Slap Música|Arrocha|Música Sertaneja (Musical Genre)|Pop|Gustavo Lima|Gus
##  $ view_count       : int  3384174 1148927 198159 228417 100811 526791 378903 2541668 401621 190898
##  $ likes            : int  49563 70255 9053 12996 188 43059 17966 190610 20538 12836 ...
##  $ dislikes         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ comment_count    : int  1297 51 19 159 2 1270 3297 22345 457 949 ...
##  $ thumbnail_link   : chr  "https://i.ytimg.com/vi/h09p0IGiKaE/default.jpg" "https://i.ytimg.com/vi/
##  $ comments_disabled: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ ratings_disabled : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ description      : chr  "Essa música faz parte do projeto PARAÍSO PARTICULAR. DVD Gravado na casa
##  $ countryName      : chr  "Brazil" "Brazil" "Brazil" "Brazil" ...
##  $ countryCode      : int  1 1 1 1 1 1 1 1 1 1 ...
```

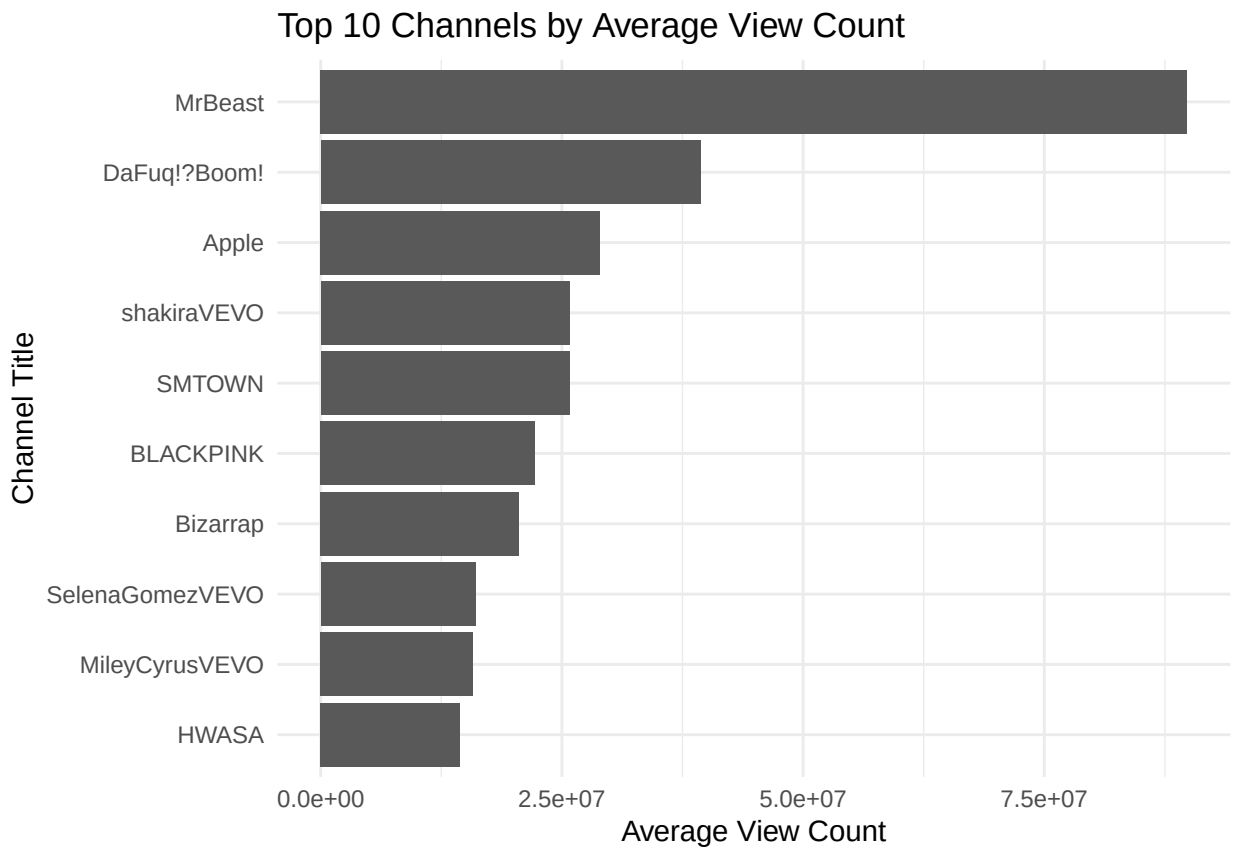
```r
library(ggplot2)
library(dplyr)
library(lubridate)

bra$publishedAt <- as.Date(bra$publishedAt)
bra %>%
  group_by(publishedAt) %>%
  summarize(AvgViewCount = mean(view_count, na.rm = TRUE)) %>%
  ggplot(aes(x = publishedAt, y = AvgViewCount)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
```

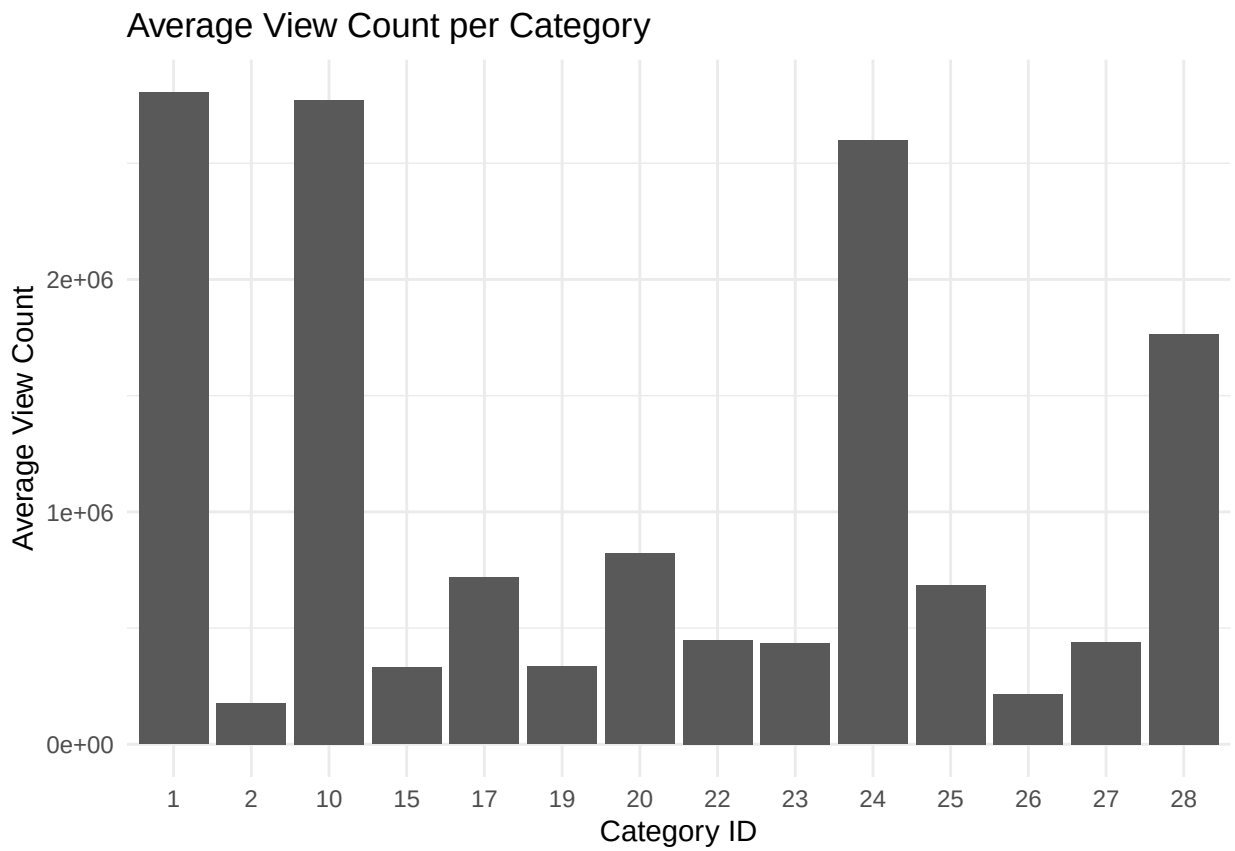
```
  labs(title = "Average View Count per Day", x = "Published Date", y = "Average View Count")
```




```
bra %>%
  group_by(channelTitle) %>%
  summarize(AvgViewCount = mean(view_count, na.rm = TRUE)) %>%
  top_n(10, AvgViewCount) %>%
  ggplot(aes(x = reorder(channelTitle, AvgViewCount), y = AvgViewCount)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  coord_flip() +
  labs(title = "Top 10 Channels by Average View Count", x = "Channel Title", y = "Average View Count")
```

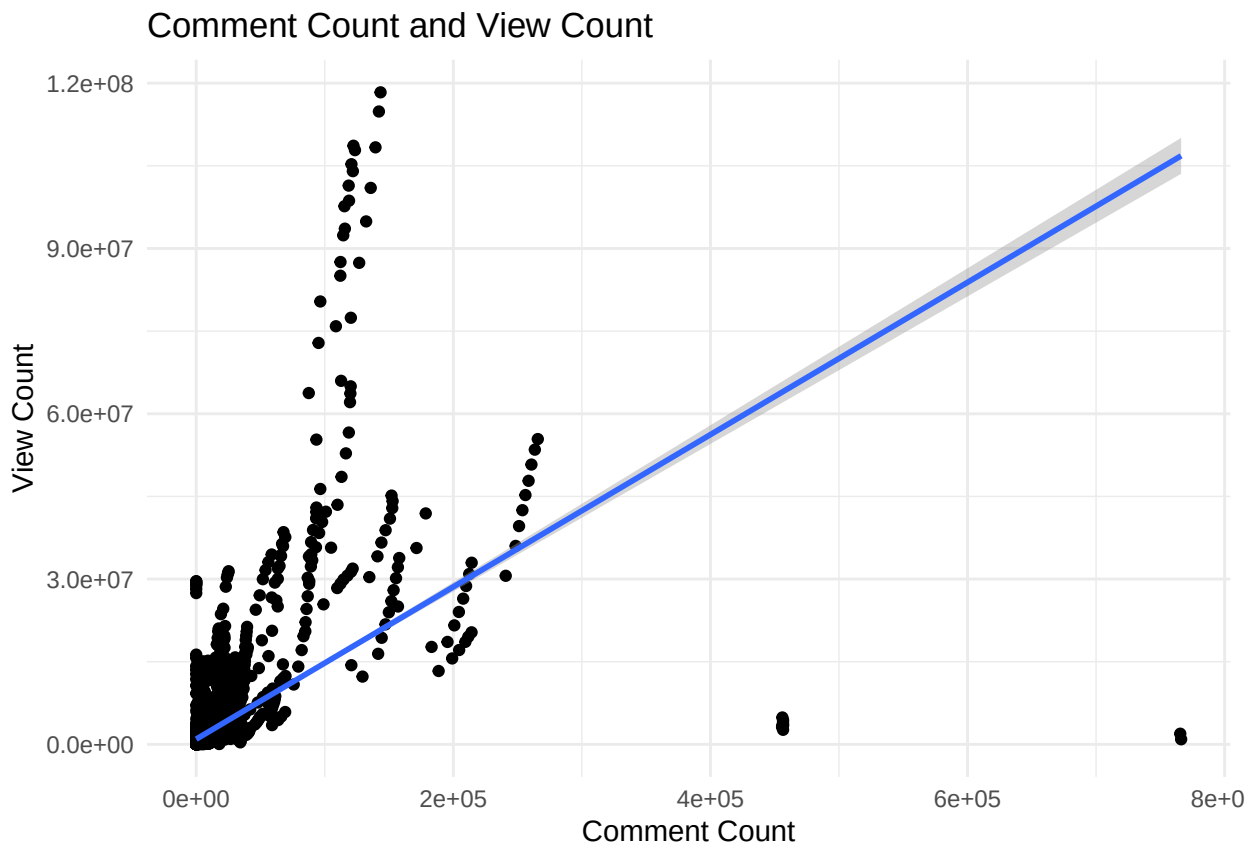



```
bra %>%
  group_by(categoryId) %>%
  summarize(AvgViewCount = mean(view_count, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(categoryId), y = AvgViewCount)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average View Count per Category", x = "Category ID", y = "Average View Count")
```




```
ggplot(bra, aes(x = comment_count, y = view_count)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_minimal() +
  labs(title = "Comment Count and View Count", x = "Comment Count", y = "View Count")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Understanding the full Multiple Linear Regression model

```
# Merging the datasets
data <- rbind(bra, usa, can, jpn, ind)

# View the structure of the merged dataset
dim(data)
```

```
## [1] 50000    18
```

```
data$countryName <- factor(data$countryName)
```

```
# Now create the linear model
full_model <- lm(view_count ~ likes + comment_count + countryCode, data = data)

# View the summary of the model
summary(full_model)
```

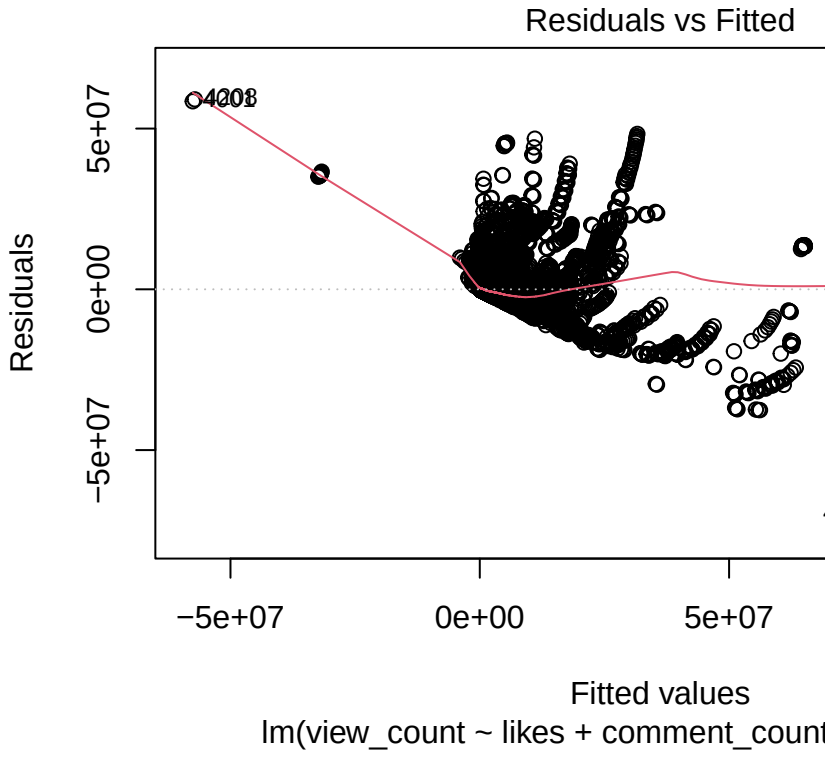
MLR model to predict final exam grades by including all prediction variables

```
##
## Call:
## lm(formula = view_count ~ likes + comment_count + countryCode,
##     data = data)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -67718333   -625715   -352186     31413  59082908
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.613e+05  3.901e+04   4.134 3.57e-05 ***
## likes          2.370e+01  7.310e-02 324.219  < 2e-16 ***
## comment_count -7.953e+01  1.052e+00 -75.626  < 2e-16 ***
## countryCode    1.212e+05  1.173e+04  10.334  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3708000 on 49996 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.8038
## F-statistic: 6.827e+04 on 3 and 49996 DF,  p-value: < 2.2e-16
```

Model Diagnostics

```
plot(full_model, which=1)
```



Checking whether variance assumption is met

Interpreting the coefficient of multiple determination R^2. We can get the value of the R^2 from the summary

```r
summary(full_model)$r.squared
```

```
## [1] 0.8037878
```

The R^2 value of 80.41% means that approximately 80.41% of the variability in view_count can be explained by the linear relationship between the response variable and the predictor variables.

```r
summary(full_model)$fstatistic
```

```
##    value    numdf    dendf
## 68269.91     3.00 49996.00
```