# Regression

9/24/22

Data set citation: https://www.kaggle.com/datasets/dmitrynikolaev/youtube-dislikes-dataset (https://www.kaggle.com/datasets/dmitrynikolaev/youtube-dislikes-dataset)

Linear Regression: Attempts to create a prediction model/line based on some predictors and some targets that are then analyzed to create a formula/line that best represents the data. Linear regression is good for data that can be linearly separated and is exceptionally easy to implement. On the other hand linear regression has high bias and can over fit the data if the data is not carefully selected.

In this block we read in the data from the file, and split it into the train and test

```
youtube <- read.csv('youtube_dislike_dataset.csv')

set.seed(1234)
i <- sample(1:nrow(youtube), nrow(youtube)*.80, replace=FALSE)
train <- youtube[i,]
test <- youtube[-i,]
```

## Five Different Data Exploration Functions with the Training Data

5 Different functions used for data exploration In this block we use different operations to both manipulate and present the data in order to have a better understanding of what the data is and how it interacts with itself.

```
str(train)
```

```
## 'data.frame':    29937 obs. of  12 variables:
##  $ video_id     : chr  "OsRHj8YEXo4" "tXZ0Qq_Nbc0" "x6k-5OIqrCc" "SiUUTejzUyk" ...
##  $ title        : chr  "DUDE BHAI CUTE BEHAN || Nishant Chaturvedi" "Jennifer Lopez | New Yea
r's Rockin' Eve Performance" "  10 .       Kranvagn   " "LAKERS at NUGGETS | FULL GAME HIGHLIGHT
S | September 24, 2020" ...
##  $ channel_id   : chr  "UCXPIEXPURo8zww575ZdcM-g" "UCr8RjWUQ_9KYcIPmWiqBroQ" "UCLZIN4aTXYm92c
1ENyN8KmA" "UCWJ2lWNubArHWmf3FIHbfcQ" ...
##  $ channel_title: chr  "Nishant Chaturvedi" "Jennifer Lopez" "Amway921" "NBA" ...
##  $ published_at : chr  "2020-10-03 08:54:51" "2021-01-01 09:00:37" "2021-09-16 04:00:14" "202
0-09-25 03:38:52" ...
##  $ view_count   : int  1393170 4987613 236109 2763214 392327 2597161 3292565 3041700 856475 6
29128 ...
##  $ likes        : int  24015 61946 18961 28069 13580 150503 113352 66013 5684 21949 ...
##  $ dislikes     : int  2507 12984 335 1262 236 726 743 1174 160 187 ...
##  $ comment_count: int  739 10909 319 6343 366 6786 4354 2653 1116 1371 ...
##  $ tags         : chr  "nishant chaturvedi. nishant chaturvedi new video. nishanr chaturvedi
comedy. nishant chaturvedi bhai behan vide"| __truncated__ " " "WOT World Of Tanks Amway921 VOD
modpack    " "NBA G League Basketball game-0041900314 game-0041900313 Jump Shot Steal Rebound As
sist Pass 2019-20 Season NBA "| __truncated__ ...
##  $ description  : chr  "World's most awaited game TAXAAL is online. If you think  you are goo
d at predictions then this game is  meant "| __truncated__ "Jennifer Lopez | New Year's Rockin'
Eve Performance\n\nHappy New Year 2021 Everyone!!!! Thanks for ringing in t"| __truncated__ "
200  \n Raid: Shadow Legends   IOS: https://clik.cc/4jyxV  Android: https://clik.cc/F5YJ5  PC: h
ttps://cli"| __truncated__ "LAKERS at NUGGETS | FULL GAME HIGHLIGHTS | September 24, 2020\n\nThe
Los Angeles Lakers defeated the Denver Nug"| __truncated__ ...
##  $ comments     : chr  "\"Life doesnt require that we be the best, only that we try our best
\" Aditi u really rocks. U can turn ny vide"| __truncated__ "When she speaks I want to cry, she
awakes so much in me with her voice. The queen J. LO People can talk bad abo"| __truncated__ "
200  \n Raid: Shadow Legends   IOS: https://clik.cc/4jyxV  Android: https://clik.cc/F5YJ5  PC: h
ttps://cli"| __truncated__ "No bs, tho. The Lakers have no choice but to close them out in the n
ext game. Cant give the Chicken Nuggets any"| __truncated__ ...
```

```
dim(train)
```

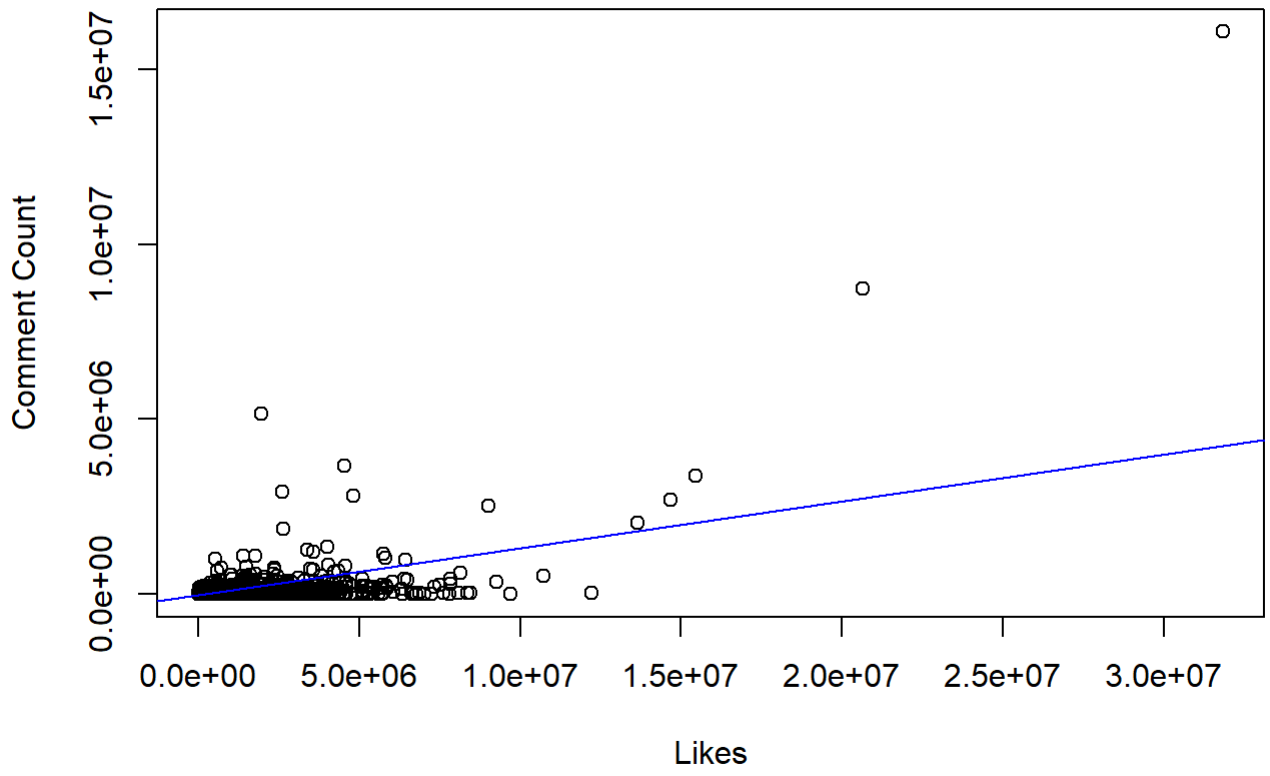
```
## [1] 29937    12
```

```
head(train, 1)
```

| | video_id | title | channel_id |
|---|---|---|---|
| | <chr> | <chr> | <chr> |
| 15241 | OsRHj8YEXo4 | DUDE BHAI CUTE BEHAN \|\| Nishant Chaturvedi | UCXPIEXPURo8zww575Zd |

1 row | 1-4 of 13 columns

```
plot(train$comment_count~train$likes , ylab = "Comment Count", xlab = "Likes")
abline(lm(train$comment_count~train$likes), col="blue")
```



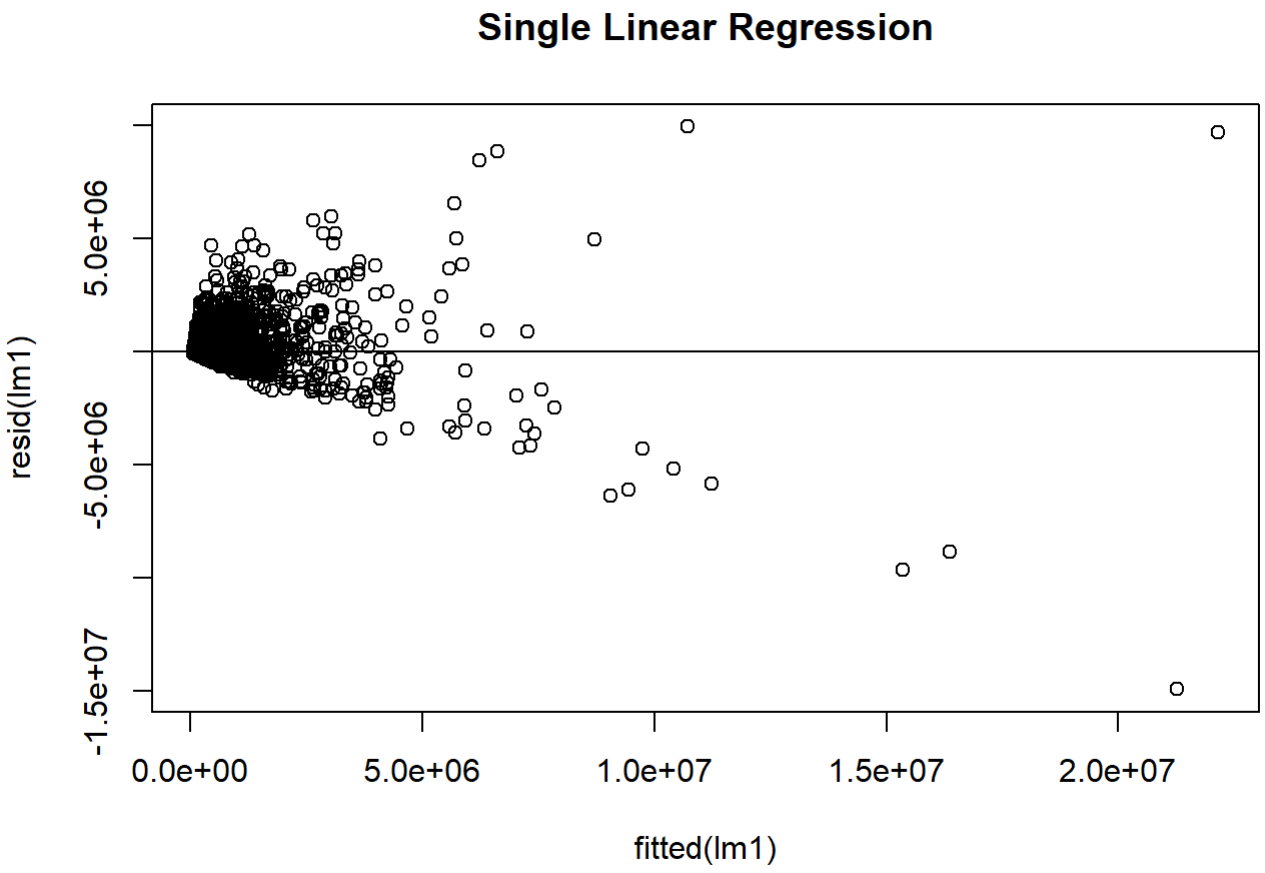

# Linear Regression model creation and printing a plot of residuals

Create the model and use it to find the predictions and the residuals as well as different stats about the line

```
lm1 <- lm(likes~view_count , data=train)
summary(lm1)
```

```
##
## Call:
## lm(formula = likes ~ view_count, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14928667    -71333    -56847     -9098   9920823
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.100e+04  1.966e+03   36.11   <2e-16 ***
## view_count  1.669e-02  7.751e-05  215.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 331500 on 29935 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.6076
## F-statistic: 4.635e+04 on 1 and 29935 DF,  p-value: < 2.2e-16
```

```
plot(fitted(lm1), resid(lm1), main = "Single Linear Regression")
abline(0,0)
```



```
pred1 <- predict(lm1, newdata=test)
cor1 <- cor(pred1, test$likes)
mse1 <- mean((pred1-test$likes)^2)
rmse1 <- sqrt(mse1)
print(paste('correlation:', cor1))
```

```
## [1] "correlation: 0.813400801558712"
```

```
print(paste('mse:', mse1))
```

```
## [1] "mse: 118163253821.982"
```

```
print(paste('rmse:', rmse1))
```

```
## [1] "rmse: 343748.823739053"
```

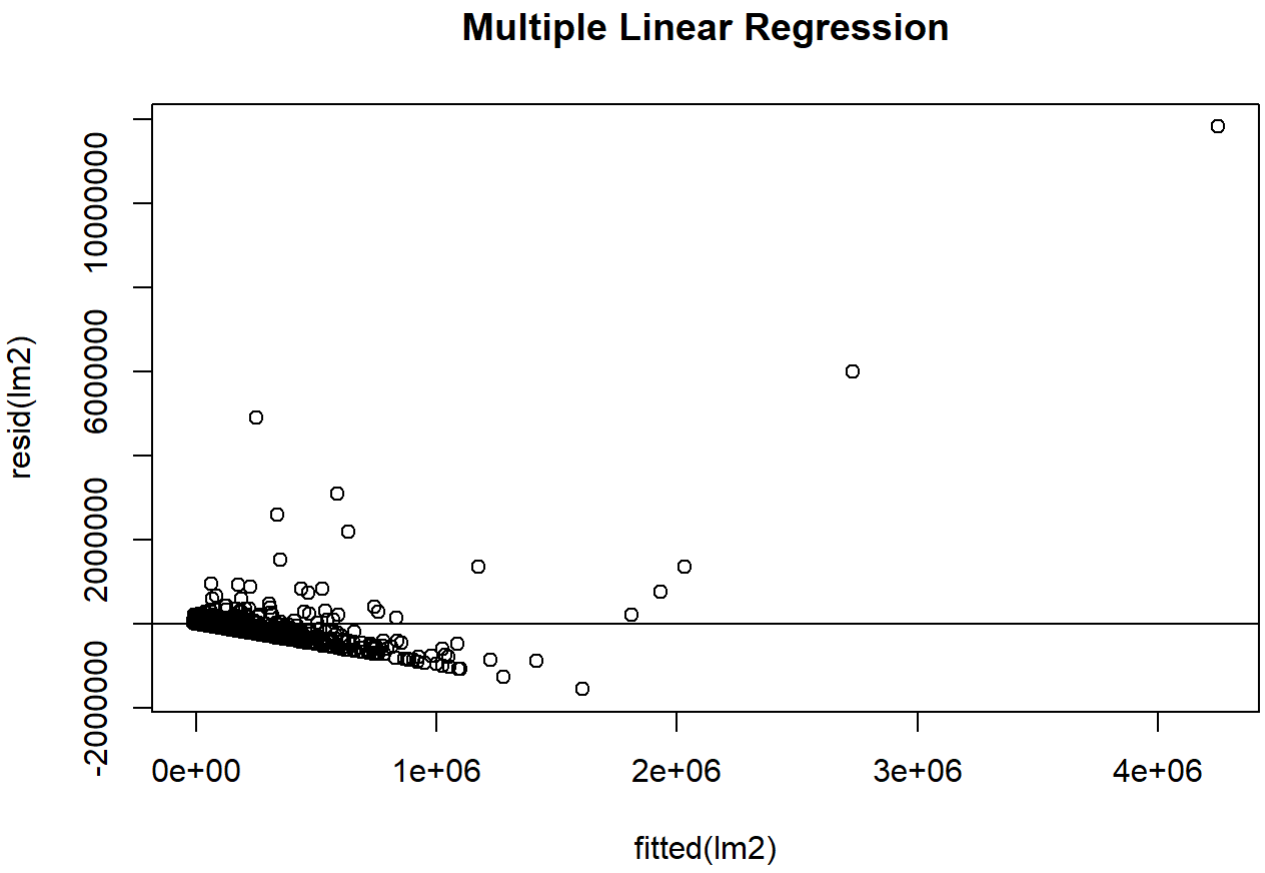
# Multiple linear regression and printing the residual plot

Using multiple linear regression we tried seeing how both the likes and view count affect the comment count of the videos

```
lm2 <- lm(comment_count~likes+view_count, data=train)
summary(lm2)
```

```
##
## Call:
## lm(formula = comment_count ~ likes + view_count, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1563089     2081     9551    11864 11821602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.221e+04  6.142e+02 -19.872   <2e-16 ***
## likes        1.301e-01  1.767e-03  73.637   <2e-16 ***
## view_count   8.923e-05  3.783e-05   2.358   0.0184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101400 on 29934 degrees of freedom
## Multiple R-squared:  0.3266, Adjusted R-squared:  0.3266
## F-statistic:  7261 on 2 and 29934 DF,  p-value: < 2.2e-16
```

```
plot(fitted(lm2), resid(lm2), main = "Multiple Linear Regression")
abline(0,0)
```




```
pred2 <- predict(lm2, newdata=test)
cor2 <- cor(pred2, test$likes+test$comment_count)
mse2 <- mean((pred1-test$likes)^2)
rmse2 <- sqrt(mse2)
print(paste('correlation:', cor2))
```

```
## [1] "correlation: 0.994764759012847"
```

```
print(paste('mse:', mse2))
```

```
## [1] "mse: 118163253821.982"
```

```
print(paste('rmse:', rmse2))
```

```
## [1] "rmse: 343748.823739053"
```

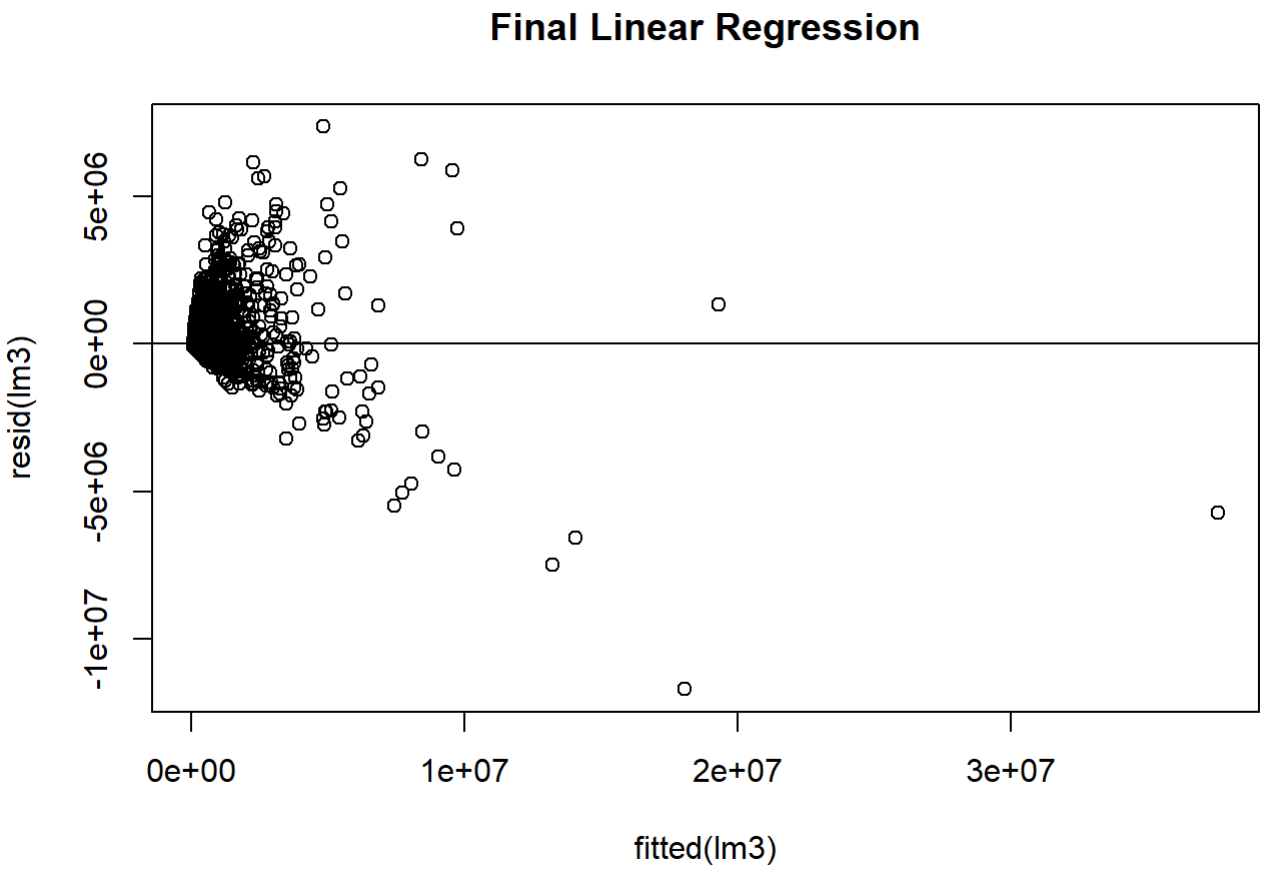
# Most accurate Linear Regression model we could create with the data

Found the highest value of R^2 that we could and all of the data that goes along with it. To do this we tried many different combinations but eventually settled on the effects that view count and comment count have on likes for a video.

```
lm3 <- lm(likes~view_count+comment_count, data=train)
summary(lm3)
```

```
##
## Call:
## lm(formula = likes ~ view_count + comment_count, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -11714670    -74315    -59650    -11606   7369158
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.449e+04  1.810e+03   41.16   <2e-16 ***
## view_count    1.402e-02  7.997e-05  175.34   <2e-16 ***
## comment_count 1.178e+00  1.600e-02   73.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 305000 on 29934 degrees of freedom
## Multiple R-squared:  0.6677, Adjusted R-squared:  0.6677
## F-statistic: 3.008e+04 on 2 and 29934 DF,  p-value: < 2.2e-16
```

```
plot(fitted(lm3), resid(lm3), main = "Final Linear Regression")
abline(0,0)
```


Final Linear Regression
resid(lm3)
5e+06
0e+00
-5e+06
-1e+07
0e+00
1e+07
2e+07
3e+07
fitted(lm3)