Youtube dataset analysis

Meirambek Ibraikhanov

Dataset overview
This dataset was collected using the YouTube API.
Main source is kaggle.com
Input variables: "category_id", "channel_title", "comment_count", "description" , "dislikes", "likes" , "publish time", "tags", "video_id", "thumbnail_link", "title", "trending_date", "video_error_or_removed", "views"
Input variables: "category_id", "category_name", "channel_id", "description", "followers", "join_date", "picture_url", "profile_url", "title", "trailer_title", "trailer_url", "videos"

Summary of dataset

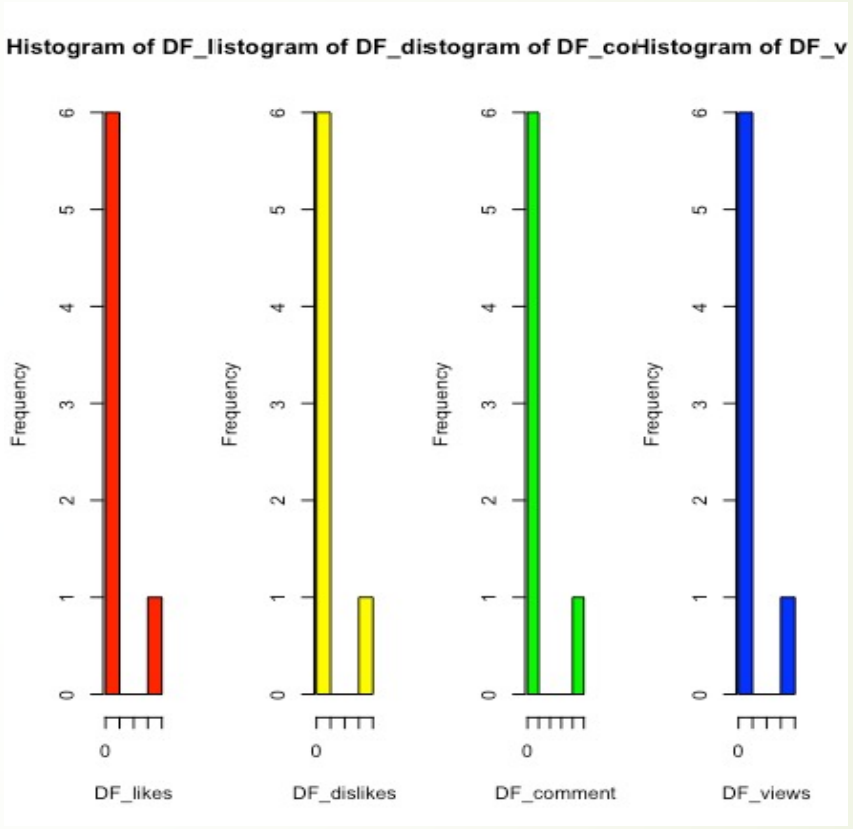
```
> summary(yd)
   video_id         trending_date         title           channel_title       category_id
 Length:81830       Length:81830       Length:81830       Length:81830       Min.   : 1
 Class :character   Class :character   Class :character   Class :character   1st Qu.:17
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :24
                                                                             Mean   :20
                                                                             3rd Qu.:24
                                                                             Max.   :43
 publish_time           tags              views              likes            dislikes
 Length:81830       Length:81830       Min.   :      549   Min.   :      0   Min.   :      0
 Class :character   Class :character   1st Qu.:   178194   1st Qu.:   3301   1st Qu.:    136
 Mode  :character   Mode  :character   Median :   496160   Median :  12834   Median :    438
                                       Mean   :  1754415   Mean   :  56939   Mean   :   2861
                                       3rd Qu.:  1355388   3rd Qu.:  40962   3rd Qu.:   1421
                                       Max.   :225211923   Max.   :5613827   Max.   :1674420
 comment_count     thumbnail_link     comments_disabled  ratings_disabled
 Min.   :      0   Length:81830       Length:81830       Length:81830
 1st Qu.:    505   Class :character   Class :character   Class :character
 Median :   1550   Mode  :character   Mode  :character   Mode  :character
 Mean   :   6746
 3rd Qu.:   4626
 Max.   :1361580
 video_error_or_removed description
 Length:81830           Length:81830
 Class :character       Class :character
 Mode  :character       Mode  :character
```

Descriptive Analysis

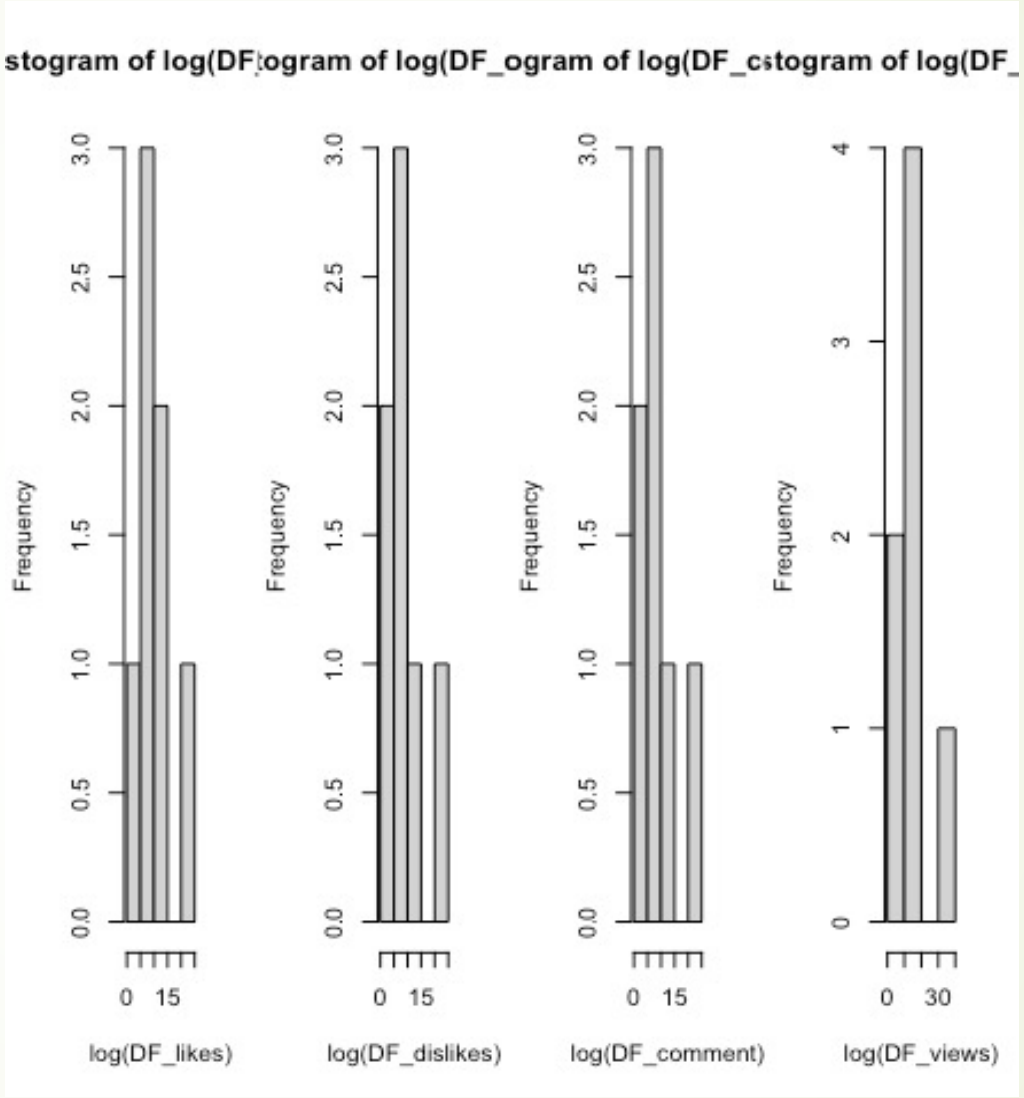
	likes	dislikes	comment_count	views
nbr.val	81830.0	81830.0	81830.0	81830.0
nbr.null	456.0	776.0	1406.0	0.0
nbr.na	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	549.0
max	5613827.0	1674420.0	1361580.0	225211923.0
range	5613827.0	1674420.0	1361580.0	225211374.0
sum	4659327076.0	234116074.0	552050013.0	143563745221.0
median	12834.0	438.0	1550.0	496160.0
mean	56939.1	2861.0	6746.3	1754414.6
SE.mean	656.9	85.8	107.0	20224.6
CI.mean.0.95	1287.5	168.2	209.7	39640.2
var	35312277731.3	602938026.2	936621538.7	33471440206069.3
std.dev	187915.6	24554.8	30604.3	5785450.7
coef.var	3.3	8.6	4.5	3.3

```
> summary(DF_likes)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
         3        972      12834 5044648195     122427 35312277731
> summary(DF_dislikes)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       9      127      438 86138020    13708 602938026
> summary(DF_comment)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       5      158     1550 133808680    18675 936621539
> summary(views)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     549   178194   496160  1754415  1355388 225211923
```

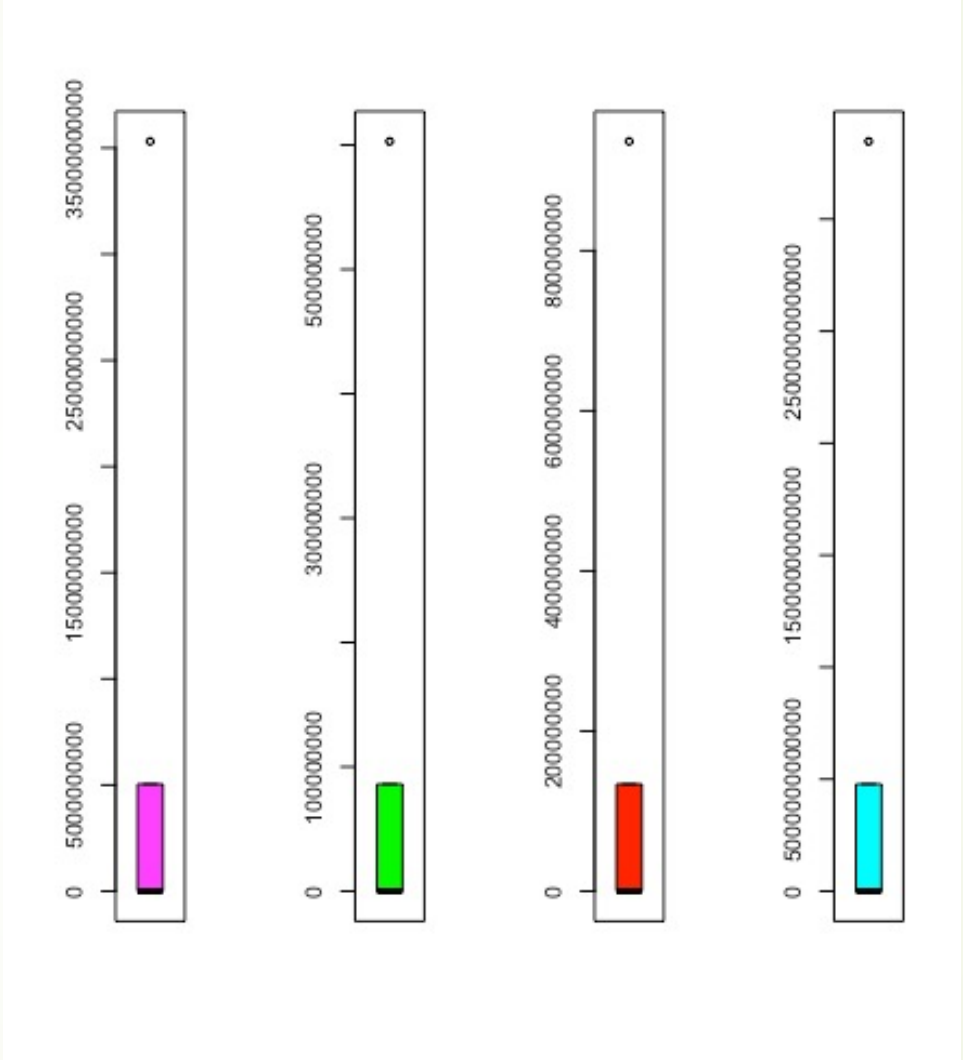
Histogram Analysis



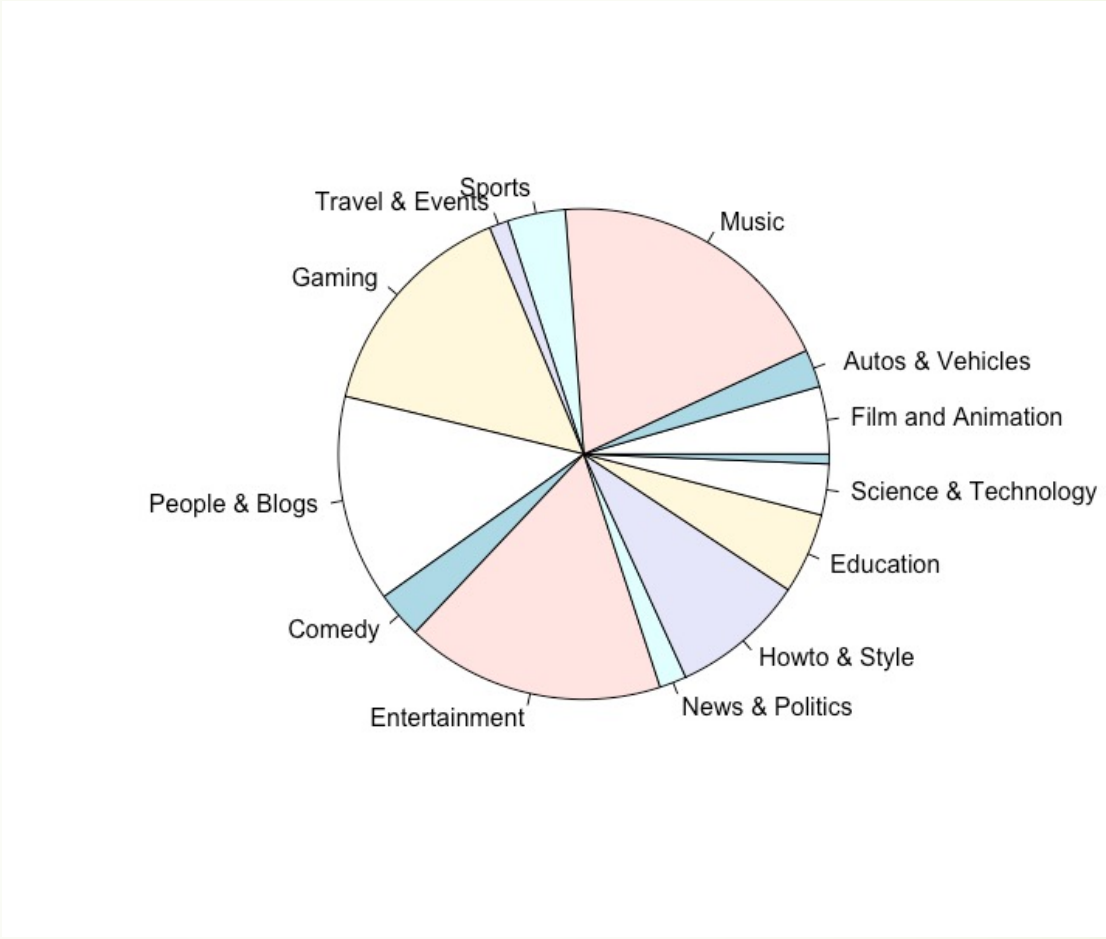
Log(histogram)



Box-plot analysis



Pie-chart Analysis



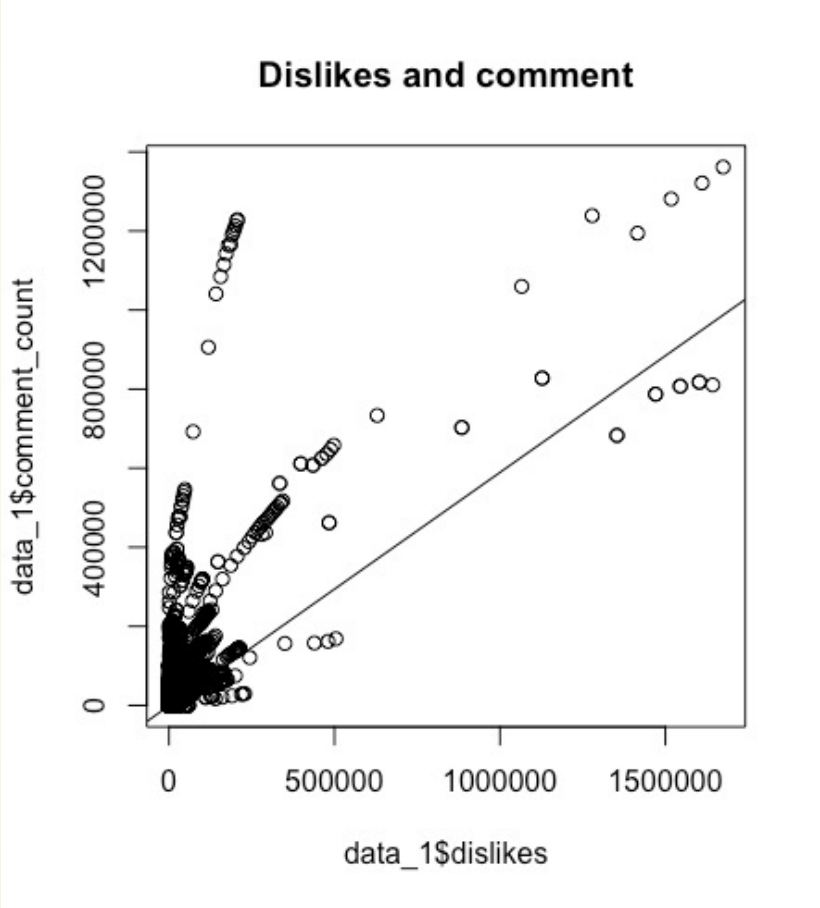
Overall mean of each variable

	likes	dislikes	comment_count	views
nbr.val	81830.0	81830.0	81830.0	81830.0
nbr.null	456.0	776.0	1406.0	0.0
nbr.na	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	549.0
max	5613827.0	1674420.0	1361580.0	225211923.0
range	5613827.0	1674420.0	1361580.0	225211374.0
sum	4659327076.0	234116074.0	552050013.0	143563745221.0
median	12834.0	438.0	1550.0	496160.0
mean	56939.1	2861.0	6746.3	1754414.6
SE.mean	656.9	85.8	107.0	20224.6
CI.mean.0.95	1287.5	168.2	209.7	39640.2
var	35312277731.3	602938026.2	936621538.7	33471440206069.3
std.dev	187915.6	24554.8	30604.3	5785450.7
coef.var	3.3	8.6	4.5	3.3

```
> stat.desc(scores, desc = F)
```

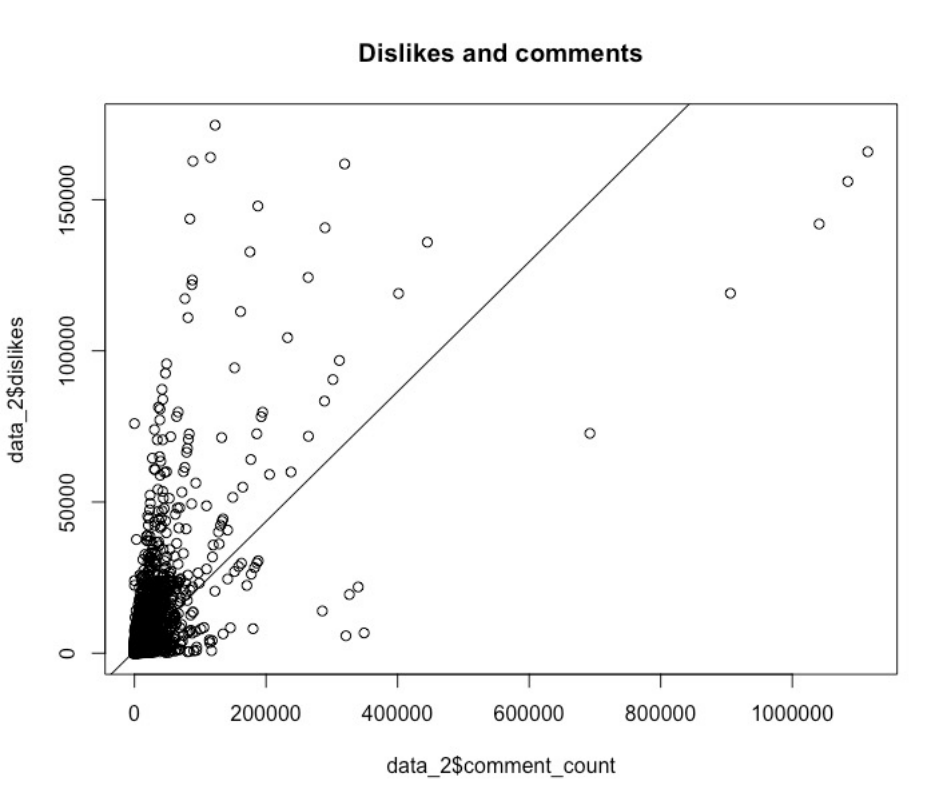
	likes	dislikes	comment_count	views
nbr.val	81830	81830	81830	81830
nbr.null	456	776	1406	0
nbr.na	0	0	0	0
min	0	0	0	549
max	5613827	1674420	1361580	225211923
range	5613827	1674420	1361580	225211374
sum	4659327076	234116074	552050013	143563745221

Linear regression of 80%



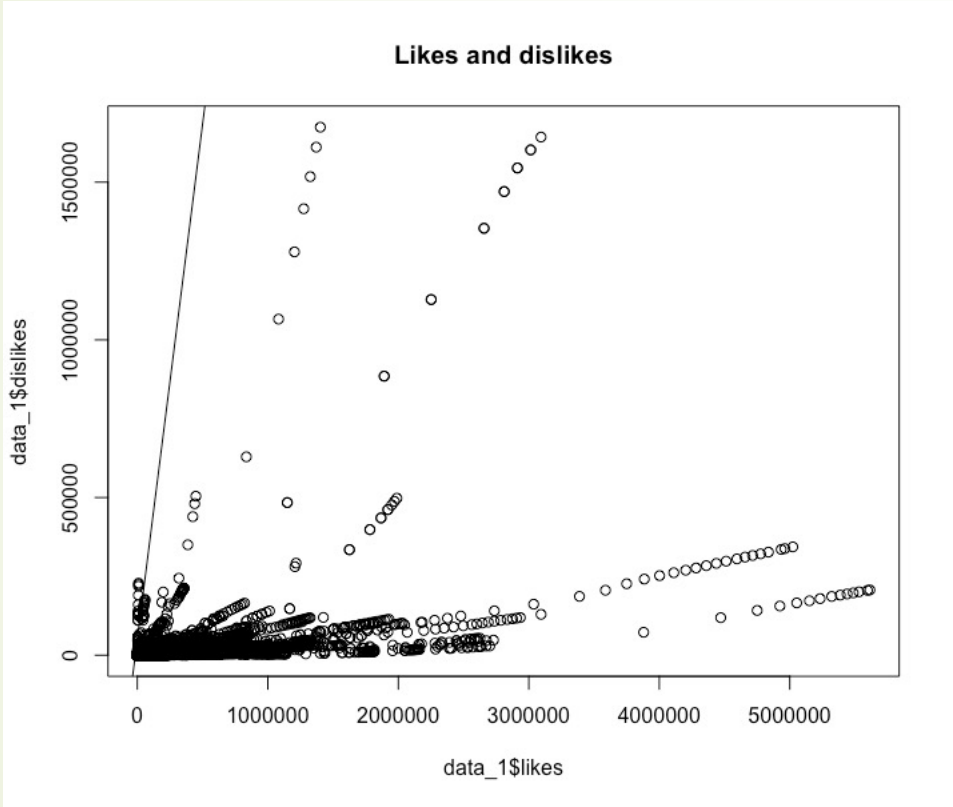

```
Call:
lm(formula = data_1$dislikes ~ data_1$comment_count)

Residuals:
    Min      1Q  Median      3Q     Max
-517194    -117     811    1082 1165657

Coefficients:
                      Estimate  Std. Error t value            Pr(>|t|)
(Intercept)          -1083.09754    77.77925   -13.9 <0.0000000000000002 ***
data_1$comment_count     0.59021     0.00235   250.7 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19400 on 65462 degrees of freedom
Multiple R-squared:  0.49,      Adjusted R-squared:  0.49
F-statistic: 6.29e+04 on 1 and 65462 DF,  p-value: <0.0000000000000002
```

Linear regression model 20%



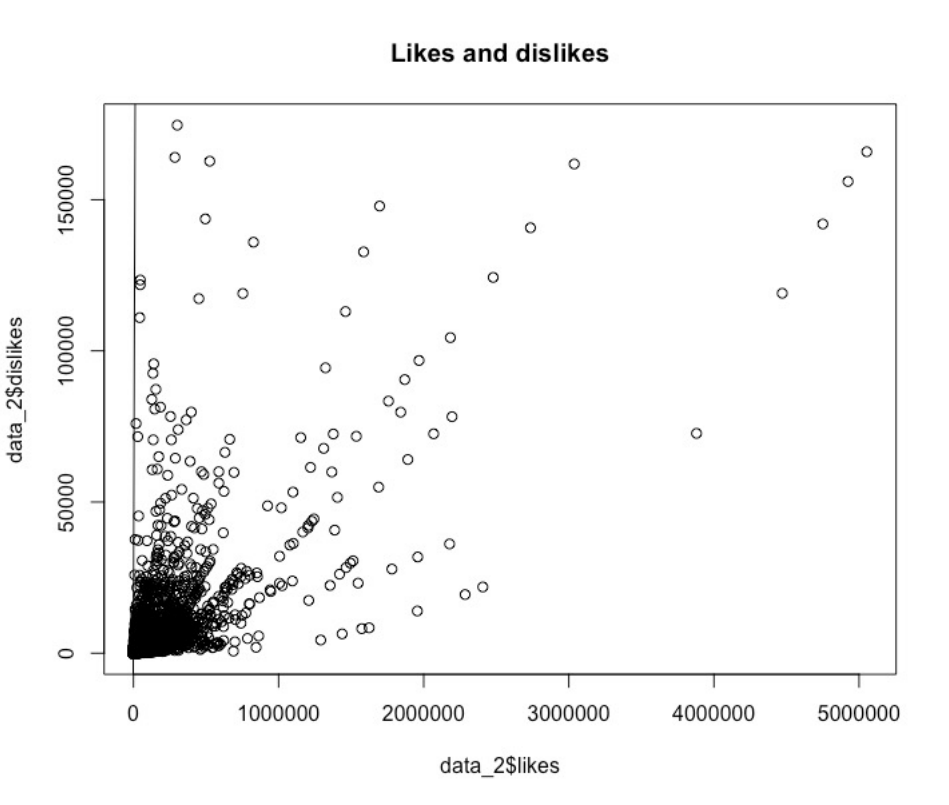

```
Call:
lm(formula = data_2$dislikes ~ data_2$comment_count)

Residuals:
   Min     1Q Median     3Q    Max
-82075   -766   -647   -371 147658

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)
(Intercept)          664.94056   43.28583    15.4 <0.0000000000000002 ***
data_2$comment_count   0.21460    0.00186   115.6 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5400 on 16365 degrees of freedom
Multiple R-squared:  0.449,     Adjusted R-squared:  0.449
F-statistic: 1.34e+04 on 1 and 16365 DF,  p-value: <0.0000000000000002
```

Linear regression of 80%



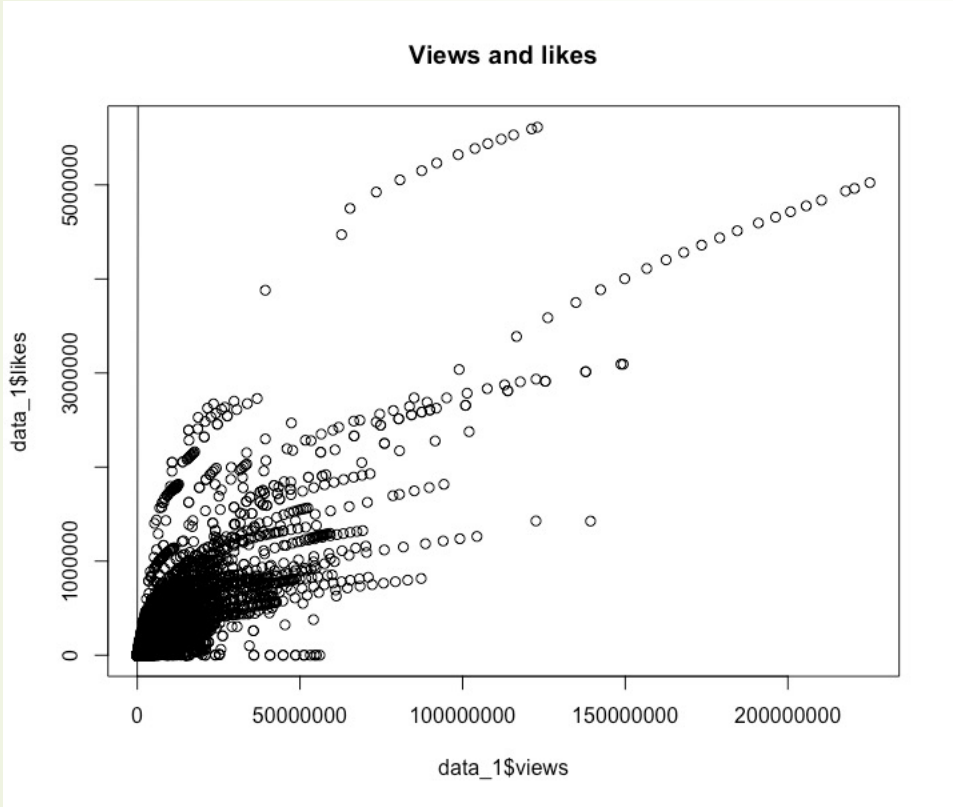

```
Call:
lm(formula = data_1$likes ~ data_1$dislikes)

Residuals:
     Min       1Q   Median       3Q      Max
-4113414   -48471   -39550   -12825  4887747

Coefficients:
                 Estimate Std. Error t value            Pr(>|t|)
(Intercept)     50798.5370   692.0957    73.4 <0.0000000000000002 ***
data_1$dislikes     3.2639     0.0253   129.1 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 176000 on 65462 degrees of freedom
Multiple R-squared:  0.203,     Adjusted R-squared:  0.203
F-statistic: 1.67e+04 on 1 and 65462 DF,  p-value: <0.0000000000000002
```

Linear regression model 20%



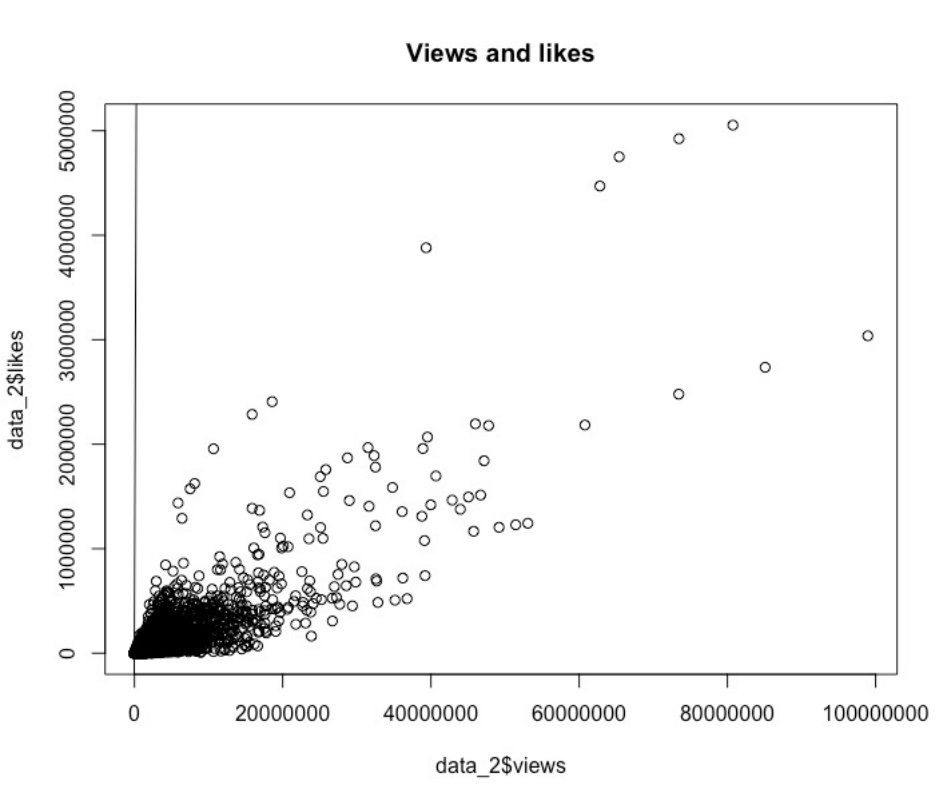

```
Call:
lm(formula = data_2$likes ~ data_2$dislikes)

Residuals:
     Min       1Q   Median       3Q      Max
-2103804   -16737   -13774     -921  2869024

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)     16359.407    843.499    19.4 <0.0000000000000002 ***
data_2$dislikes    13.681      0.113   121.5 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105000 on 16365 degrees of freedom
Multiple R-squared:  0.474,     Adjusted R-squared:  0.474
F-statistic: 1.48e+04 on 1 and 16365 DF,  p-value: <0.0000000000000002
```

Linear regression of 80%




```
Call:
lm(formula = data_1$views ~ data_1$likes)

Residuals:
      Min        1Q    Median        3Q       Max
-64998045   -355971   -207643     79897 100834184

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)  259250.846  13618.759      19 <0.0000000000000002 ***
data_1$likes     26.826      0.066     406 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3330000 on 65462 degrees of freedom
Multiple R-squared:  0.716,     Adjusted R-squared:  0.716
F-statistic: 1.65e+05 on 1 and 65462 DF,  p-value: <0.0000000000000002
```

Linear regression model 20%



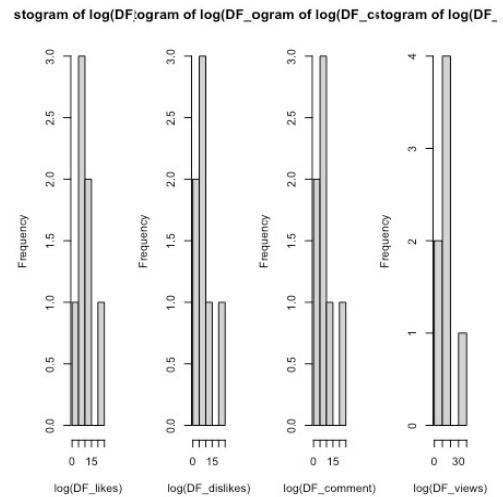

```
Call:
lm(formula = data_2$views ~ data_2$likes)

Residuals:
      Min        1Q    Median        3Q       Max
-35627925   -388457   -275735     10707  40157940

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  406455.7715  14383.6356    28.3 <0.0000000000000002 ***
data_2$likes     19.2191      0.0958   200.7 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1770000 on 16365 degrees of freedom
Multiple R-squared:  0.711,     Adjusted R-squared:  0.711
F-statistic: 4.03e+04 on 1 and 16365 DF,  p-value: <0.0000000000000002
```

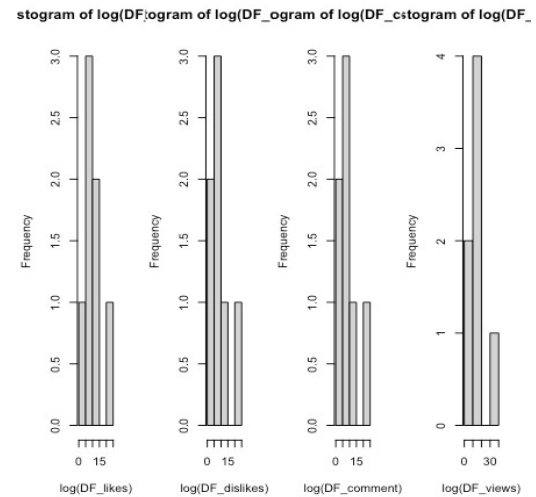
I would say model 3 is better that two models
Relationship between views and likes stronger that other models
This model help us understand relationship between views and likes

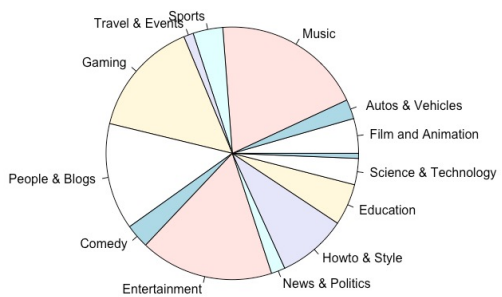
Packages	
pastecs	It allows to transform an irregular time series into a regular one, and to analyze and decompose regular time series.
ggplot2	It is a system for declaratively creating graphics, based on The Grammar of Graphics.
dplyr	It is a grammar of data manipulation , providing a consistent set of verbs that help you solve the most common data manipulation challenges.
Stat.desc	Part of the pastecs package

```
> summary(yd)
   video_id         trending_date        title           channel_title      category_id
 Length:81830       Length:81830       Length:81830       Length:81830       Min.   : 1
 Class :character   Class :character   Class :character   Class :character   1st Qu.:17
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :24
                                                                             Mean   :20
                                                                             3rd Qu.:24
                                                                             Max.   :43
 publish_time           tags              views              likes            dislikes
 Length:81830       Length:81830       Min.   :      549   Min.   :     0   Min.   :     0
 Class :character   Class :character   1st Qu.:   178194   1st Qu.:  3301   1st Qu.:   136
 Mode  :character   Mode  :character   Median :   496160   Median :  12834   Median :   438
                                       Mean   :  1754415   Mean   :  56939   Mean   :  2861
                                       3rd Qu.:  1355388   3rd Qu.:  40962   3rd Qu.:  1421
                                       Max.   :225211923   Max.   :5613827   Max.   :1674420
 comment_count     thumbnail_link     comments_disabled  ratings_disabled
 Min.   :     0   Length:81830       Length:81830       Length:81830
 1st Qu.:   505   Class :character   Class :character   Class :character
 Median :  1550   Mode  :character   Mode  :character   Mode  :character
 Mean   :  6746
 3rd Qu.:  4626
 Max.   :1361580
 video_error_or_removed description
 Length:81830           Length:81830
 Class :character       Class :character
 Mode  :character       Mode  :character
```

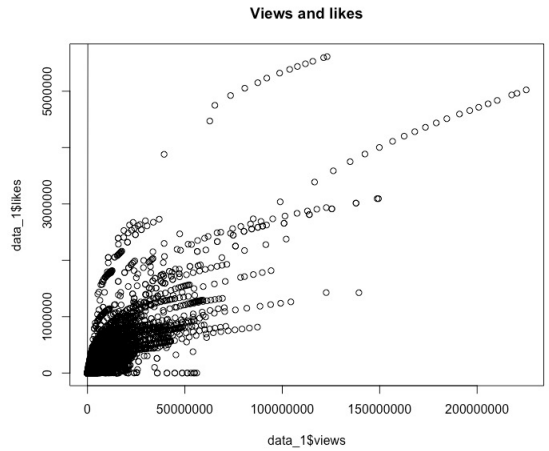
	likes	dislikes	comment_count	views
nbr.val	81830.0	81830.0	81830.0	81830.0
nbr.null	456.0	776.0	1406.0	0.0
nbr.na	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	549.0
max	5613827.0	1674420.0	1361580.0	225211923.0
range	5613827.0	1674420.0	1361580.0	225211374.0
sum	4659327076.0	234116074.0	552050013.0	143563745221.0
median	12834.0	438.0	1550.0	496160.0
mean	56939.1	2861.0	6746.3	1754414.6
SE.mean	656.9	85.8	107.0	20224.6
CI.mean.0.95	1287.5	168.2	209.7	39640.2
var	35312277731.3	602938026.2	936621538.7	33471440206069.3
std.dev	187915.6	24554.8	30604.3	5785450.7
coef.var	3.3	8.6	4.5	3.3







	likes	dislikes	comment_count	views
nbr.val	81830.0	81830.0	81830.0	81830.0
nbr.null	456.0	776.0	1406.0	0.0
nbr.na	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	549.0
max	5613827.0	1674420.0	1361580.0	225211923.0
range	5613827.0	1674420.0	1361580.0	225211374.0
sum	4659327076.0	234116074.0	552050013.0	143563745221.0
median	12834.0	438.0	1550.0	496160.0
mean	56939.1	2861.0	6746.3	1754414.6
SE.mean	656.9	85.8	107.0	20224.6
CI.mean.0.95	1287.5	168.2	209.7	39640.2
var	35312277731.3	602938026.2	936621538.7	33471440206069.3
std.dev	187915.6	24554.8	30604.3	5785450.7
coef.var	3.3	8.6	4.5	3.3




```
Call:
lm(formula = data_1$views ~ data_1$likes)

Residuals:
      Min        1Q    Median        3Q       Max
-64998045   -355971   -207643     79897 100834184

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  259250.846  13618.759      19 <0.0000000000000002 ***
data_1$likes      26.826      0.066     406 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3330000 on 65462 degrees of freedom
Multiple R-squared:  0.716,     Adjusted R-squared:  0.716
F-statistic: 1.65e+05 on 1 and 65462 DF,  p-value: <0.0000000000000002
```