R_T3_master

2024-09-17

Load required packages

```
# Install necessary packages
if(!require(tidyverse)) install.packages('tidyverse')
```

```
## Loading required package: tidyverse

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

```
if(!require(ggplot2)) install.packages('ggplot2')
if(!require(lmtest)) install.packages('lmtest')
```

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

```
if(!require(broom)) install.packages('broom')
```

```
## Loading required package: broom
```

```
if(!require(readr)) install.packages('readr')
if(!require(dplyr)) install.packages('dplyr')
```

Load the dataset

```r
df <- read_csv('data.csv')
```

```
## New names:
## Rows: 2550 Columns: 27
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (27): ...1, Talk_ID, Funny, Beautiful, Ingenious, Courageous, Longwinded...
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

```r
head(df)
```

```
## # A tibble: 6 x 27
##    ...1 Talk_ID Funny Beautiful Ingenious Courageous Longwinded Confusing
##   <dbl>   <dbl> <dbl>     <dbl>     <dbl>      <dbl>      <dbl>     <dbl>
## 1     0       1 20.9       4.87      6.47       3.47      0.412     0.258
## 2     1       2 18.5       1.97      1.91       4.73      3.85      2.11
## 3     2       3 34.1       2.12      6.47       1.59      2.76      0.955
## 4     3       4  1.58      7.80      2.81      20.4       1.42      0.857
## 5     4       5  5.42      3.68     12.5        1.24      0.429     0.281
## 6     5       6  7.17      4.60      2.58       4.69      1.86      1.96
## # i 19 more variables: Informative <dbl>, Fascinating <dbl>,
## #   Unconvincing <dbl>, Persuasive <dbl>, Jaw_dropping <dbl>, OK <dbl>,
## #   Obnoxious <dbl>, Inspiring <dbl>, total <dbl>, log_views <dbl>,
## #   Other_TED <dbl>, TED <dbl>, TEDGlobal <dbl>, TEDMED <dbl>, TEDSalon <dbl>,
## #   TEDWomen <dbl>, TEDx <dbl>, duration <dbl>, months_ago <dbl>
```

Data Cleaning: Drop unnecessary columns

```r
df <- df %>% select(-c('Talk_ID', 'total', 'Beautiful', 'Ingenious'))
head(df)
```

```
## # A tibble: 6 x 23
##    ...1 Funny Courageous Longwinded Confusing Informative Fascinating
##   <dbl> <dbl>      <dbl>      <dbl>     <dbl>       <dbl>       <dbl>
## 1     0 20.9         3.47      0.412     0.258        7.83       11.3
## 2     1 18.5         4.73      3.85      2.11        15.1         4.49
## 3     2 34.1         1.59      2.76      0.955       14.0         5.87
## 4     3  1.58       20.4       1.42      0.857       10.2         3.54
## 5     4  5.42        1.24      0.429     0.281       21.2        18.0
## 6     5  7.17        4.69      1.86      1.96         6.76        8.79
## # i 16 more variables: Unconvincing <dbl>, Persuasive <dbl>,
## #   Jaw_dropping <dbl>, OK <dbl>, Obnoxious <dbl>, Inspiring <dbl>,
## #   log_views <dbl>, Other_TED <dbl>, TED <dbl>, TEDGlobal <dbl>, TEDMED <dbl>,
## #   TEDSalon <dbl>, TEDWomen <dbl>, TEDx <dbl>, duration <dbl>,
## #   months_ago <dbl>
```

```r
colnames(df)
```

```
##  [1] "...1"          "Funny"         "Courageous"    "Longwinded"    "Confusing"    
##  [6] "Informative"   "Fascinating"   "Unconvincing"  "Persuasive"    "Jaw_dropping" 
## [11] "OK"            "Obnoxious"     "Inspiring"     "log_views"     "Other_TED"    
## [16] "TED"           "TEDGlobal"     "TEDMED"        "TEDSalon"      "TEDWomen"     
## [21] "TEDx"          "duration"      "months_ago"
```

Feature Engineering: Creating new features

```r
df <- df %>%
  mutate(ted_33 = TED^3,
         ted_global_33 = TEDGlobal^3,
         ted_x_33 = TEDx^3)
```

Linear Regression Model

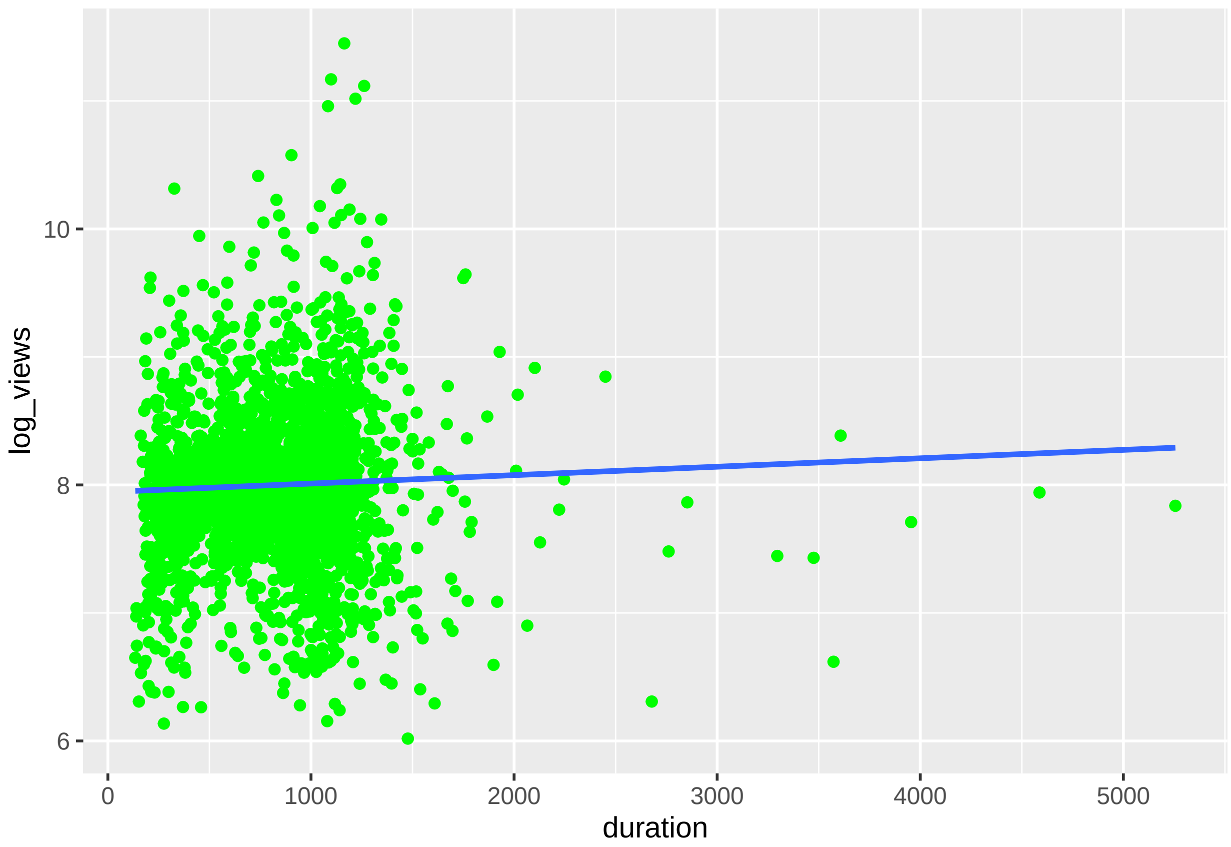
```r
formula <- as.formula('log_views ~ .')
model <- lm(formula, data = df)
summary(model)
```

```
## 
## Call:
## lm(formula = formula, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.38982 -0.14544 -0.01048  0.13100  2.96197 
## 
## Coefficients: (4 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    5.714e+00  1.317e-01  43.386  < 2e-16 ***
## ...1           8.259e-04  4.579e-05  18.036  < 2e-16 ***
## Funny          3.356e-02  1.037e-03  32.352  < 2e-16 ***
## Courageous     1.289e-02  1.656e-03   7.787 9.95e-15 ***
## Longwinded    -1.489e-02  5.930e-03  -2.512 0.012077 *  
## Confusing      4.222e-03  7.896e-03   0.535 0.592885    
## Informative    8.947e-03  1.413e-03   6.334 2.82e-10 ***
## Fascinating    4.147e-02  1.709e-03  24.267  < 2e-16 ***
## Unconvincing   3.559e-02  3.855e-03   9.232  < 2e-16 ***
## Persuasive     3.728e-02  1.744e-03  21.383  < 2e-16 ***
## Jaw_dropping   4.385e-02  1.653e-03  26.526  < 2e-16 ***
## OK            -2.659e-03  4.573e-03  -0.581 0.561029    
## Obnoxious      1.622e-02  5.674e-03   2.859 0.004289 ** 
## Inspiring      3.611e-02  9.946e-04  36.309  < 2e-16 ***
## Other_TED     -6.277e-02  2.204e-02  -2.848 0.004430 ** 
## TED           -4.125e-02  1.842e-02  -2.240 0.025164 *  
## TEDGlobal     -2.531e-02  2.132e-02  -1.187 0.235267
## TEDMED        -2.609e-02  4.065e-02  -0.642 0.521019
## TEDSalon      -4.096e-02  3.302e-02  -1.240 0.214976
## TEDWomen      -5.129e-02  3.537e-02  -1.450 0.147225
## TEDx                  NA         NA      NA       NA
## duration      -7.717e-05  2.039e-05  -3.784 0.000158 ***
## months_ago     1.475e-04  8.951e-04   0.165 0.869100
## ted_33                NA         NA      NA       NA
## ted_global_33         NA         NA      NA       NA
## ted_x_33              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3124 on 2528 degrees of freedom
## Multiple R-squared:  0.7344, Adjusted R-squared:  0.7322
## F-statistic: 332.8 on 21 and 2528 DF,  p-value: < 2.2e-16
```

Visualize: Scatter plot and regression line

```r
# Visualize: Scatter plot and regression line
ggplot(df, aes(x = duration, y = log_views)) +
  geom_point(color = 'green') +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(x = 'duration', y = 'log_views')
```

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



# Model Evaluation: MSE and R-squared

```
predictions <- predict(model, df)
mse <- mean((df$log_views - predictions)^2)
r2 <- summary(model)$r.squared

mse
```

```
## [1] 0.09672371
```

```
r2
```

```
## [1] 0.7343872
```