Final Project Summary Stats & Plots

Julian Murillo

11/12/2019

MASHABLE DATA... ONLINE MEDIA POPULARITY

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(leaps)

# Read-in Data and Load Libraries into current session
raw_data <- read.csv("~/Desktop/MGSC 310/MGSC 310 Final Project/OnlineNewsPopularityData.csv")
sapply(raw_data,is.numeric)
```

```
##                           url                     timedelta
##                         FALSE                          TRUE
##                n_tokens_title               n_tokens_content
##                          TRUE                          TRUE
##               n_unique_tokens               n_non_stop_words
##                          TRUE                          TRUE
##      n_non_stop_unique_tokens                     num_hrefs
##                          TRUE                          TRUE
##                num_self_hrefs                      num_imgs
##                          TRUE                          TRUE
##                    num_videos           average_token_length
##                          TRUE                          TRUE
##                  num_keywords      data_channel_is_lifestyle
##                          TRUE                          TRUE
## data_channel_is_entertainment            data_channel_is_bus
##                          TRUE                          TRUE
##        data_channel_is_socmed           data_channel_is_tech
##                          TRUE                          TRUE
##         data_channel_is_world                     kw_min_min
##                          TRUE                          TRUE
##                    kw_max_min                     kw_avg_min
##                          TRUE                          TRUE
##                    kw_min_max                     kw_max_max
##                          TRUE                          TRUE
##                    kw_avg_max                     kw_min_avg
##                          TRUE                          TRUE
##                    kw_max_avg                     kw_avg_avg
##                          TRUE                          TRUE
```

```
##     self_reference_min_shares     self_reference_max_shares
##                          TRUE                          TRUE
##    self_reference_avg_sharess            weekday_is_monday
##                          TRUE                          TRUE
##           weekday_is_tuesday          weekday_is_wednesday
##                          TRUE                          TRUE
##          weekday_is_thursday            weekday_is_friday
##                          TRUE                          TRUE
##          weekday_is_saturday            weekday_is_sunday
##                          TRUE                          TRUE
##                    is_weekend                        LDA_00
##                          TRUE                          TRUE
##                        LDA_01                        LDA_02
##                          TRUE                          TRUE
##                        LDA_03                        LDA_04
##                          TRUE                          TRUE
##          global_subjectivity     global_sentiment_polarity
##                          TRUE                          TRUE
##    global_rate_positive_words    global_rate_negative_words
##                          TRUE                          TRUE
##          rate_positive_words           rate_negative_words
##                          TRUE                          TRUE
##         avg_positive_polarity         min_positive_polarity
##                          TRUE                          TRUE
##         max_positive_polarity         avg_negative_polarity
##                          TRUE                          TRUE
##         min_negative_polarity         max_negative_polarity
##                          TRUE                          TRUE
##           title_subjectivity      title_sentiment_polarity
##                          TRUE                          TRUE
##        abs_title_subjectivity  abs_title_sentiment_polarity
##                          TRUE                          TRUE
##                        shares
##                          TRUE
```

REMOVE NON-PREDICTIVE VARIABLES FOR VAR SELECTION

```
raw_data <- select(raw_data,-c(url,timedelta))
summary(raw_data$shares)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     946    1400    3395    2800  843300
```

```
raw_data <- raw_data %>% mutate(popular_article = ifelse(shares > 1400,1,0))

#Take out "shares"...so we can iterate over whole data set without any problems
binomial_data <- select(raw_data, -c(shares))
```

Here we decided to apply our variable selection methods using the binary/binomial predictor "popular_article". This is due to the fact that when we tried to run our variable selection methods predicting actual count/number of shares, often times the best methods (ridge, lasso, elastic-net) did not select any of our 58 predictor variables. We figured that this must be because of how difficult it would actually be to predict the exact number of shares for any given web article. We re-ran our variable-selection tests using our binomial training set which contains only binomial target ('popular_article') and all of our 58 potential predictors.

TEST AND TRAINING SPLIT

```
set.seed(1861)
train_idx <- sample(1:nrow(binomial_data), size = floor(nrow(binomial_data) * .75))
binomial_train <- binomial_data %>% slice(train_idx)
binomial_test <- binomial_data %>% slice(-train_idx)
#we have 58 predictors...1 target variable (shares: converted to binary)
```

CORRELATION MATRIX TO DETECT COLINEARITY

```
cor_matrix <- cor(binomial_train)
cor_matrix <- round(cor_matrix, 2)

#Variance Inflation Factor (VIF)...anything over like 5 or 10 is problematic (regular OLS model)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
lm_mod1 <- lm(shares ~.,
           data = raw_data)
summary(lm_mod1)
```

```
##
## Call:
## lm(formula = shares ~ ., data = raw_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21443  -2692   -488    646 835436
##
## Coefficients: (2 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.001e+06  6.018e+06  -0.166 0.867855
## n_tokens_title                 8.823e+01  2.815e+01   3.135 0.001722 **
## n_tokens_content               4.597e-01  2.195e-01   2.095 0.036216 *
## n_unique_tokens                3.974e+03  1.884e+03   2.110 0.034883 *
## n_non_stop_words              -1.124e+03  5.803e+03  -0.194 0.846452
## n_non_stop_unique_tokens      -9.434e+02  1.600e+03  -0.590 0.555387
## num_hrefs                      1.858e+01  6.587e+00   2.820 0.004800 **
## num_self_hrefs                -3.434e+01  1.750e+01  -1.962 0.049785 *
## num_imgs                       9.923e+00  8.779e+00   1.130 0.258336
## num_videos                     3.852e+00  1.546e+01   0.249 0.803284
## average_token_length          -4.652e+02  2.384e+02  -1.951 0.051054 .
## num_keywords                   4.524e+00  3.648e+01   0.124 0.901293
## data_channel_is_lifestyle     -8.668e+02  3.875e+02  -2.237 0.025284 *
## data_channel_is_entertainment -7.794e+02  2.508e+02  -3.108 0.001884 **
## data_channel_is_bus           -4.614e+02  3.758e+02  -1.228 0.219610
## data_channel_is_socmed        -1.321e+03  3.661e+02  -3.610 0.000307 ***
## data_channel_is_tech          -9.927e+02  3.648e+02  -2.721 0.006509 **
## data_channel_is_world         -3.989e+02  3.695e+02  -1.080 0.280326
```

```
## kw_min_min                    4.199e-01  1.594e+00   0.263 0.792304
## kw_max_min                    4.479e-02  4.922e-02   0.910 0.362844
## kw_avg_min                   -6.119e-02  3.023e-01  -0.202 0.839593
## kw_min_max                   -1.391e-03  1.152e-03  -1.207 0.227374
## kw_max_max                   -1.607e-04  5.681e-04  -0.283 0.777358
## kw_avg_max                   -1.609e-04  8.143e-04  -0.198 0.843402
## kw_min_avg                   -2.937e-01  7.431e-02  -3.952 7.75e-05 ***
## kw_max_avg                   -1.133e-01  2.494e-02  -4.541 5.61e-06 ***
## kw_avg_avg                    9.789e-01  1.423e-01   6.878 6.14e-12 ***
## self_reference_min_shares     2.503e-02  7.388e-03   3.387 0.000706 ***
## self_reference_max_shares     5.375e-03  4.009e-03   1.341 0.179997
## self_reference_avg_sharess   -7.820e-03  1.025e-02  -0.763 0.445484
## weekday_is_monday             9.505e+02  2.589e+02   3.671 0.000242 ***
## weekday_is_tuesday            5.495e+02  2.554e+02   2.152 0.031442 *
## weekday_is_wednesday          7.142e+02  2.553e+02   2.797 0.005158 **
## weekday_is_thursday           4.821e+02  2.558e+02   1.885 0.059441 .
## weekday_is_friday             3.362e+02  2.645e+02   1.271 0.203829
## weekday_is_saturday           1.930e+02  3.147e+02   0.613 0.539662
## weekday_is_sunday                    NA         NA      NA       NA
## is_weekend                           NA         NA      NA       NA
## LDA_00                        9.990e+05  6.018e+06   0.166 0.868160
## LDA_01                        9.992e+05  6.018e+06   0.166 0.868122
## LDA_02                        9.990e+05  6.018e+06   0.166 0.868153
## LDA_03                        9.996e+05  6.018e+06   0.166 0.868072
## LDA_04                        9.991e+05  6.018e+06   0.166 0.868135
## global_subjectivity           1.426e+03  8.355e+02   1.707 0.087774 .
## global_sentiment_polarity     6.868e+02  1.637e+03   0.419 0.674856
## global_rate_positive_words   -1.026e+04  7.035e+03  -1.458 0.144838
## global_rate_negative_words   -1.321e+03  1.343e+04  -0.098 0.921608
## rate_positive_words           1.023e+03  5.671e+03   0.180 0.856829
## rate_negative_words           1.369e+03  5.716e+03   0.240 0.810656
## avg_positive_polarity        -1.257e+03  1.342e+03  -0.937 0.348916
## min_positive_polarity        -1.534e+03  1.123e+03  -1.365 0.172200
## max_positive_polarity         3.184e+02  4.232e+02   0.752 0.451848
## avg_negative_polarity        -1.601e+03  1.236e+03  -1.296 0.194970
## min_negative_polarity         8.613e+01  4.505e+02   0.191 0.848396
## max_negative_polarity        -2.547e+02  1.027e+03  -0.248 0.804189
## title_subjectivity           -2.691e+02  2.692e+02  -1.000 0.317454
## title_sentiment_polarity     -1.419e+01  2.459e+02  -0.058 0.954005
## abs_title_subjectivity        3.386e+02  3.575e+02   0.947 0.343588
## abs_title_sentiment_polarity  6.505e+02  3.885e+02   1.674 0.094091 .
## popular_article               4.665e+03  1.210e+02  38.557  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11290 on 39586 degrees of freedom
## Multiple R-squared:  0.05846,    Adjusted R-squared:  0.0571
## F-statistic: 43.12 on 57 and 39586 DF,  p-value: < 2.2e-16
```

```
VIF_1 <- ols_vif_tol(lm_mod1)
VIF_1
```

```
## # A tibble: 59 x 3
##    Variables                Tolerance       VIF
##    <chr>                        <dbl>     <dbl>
```

```
##  1 n_tokens_title           0.908           1.10
##  2 n_tokens_content         0.301           3.32
##  3 n_unique_tokens          0.0000731   13677.
##  4 n_non_stop_words         0.00000349 286608.
##  5 n_non_stop_unique_tokens 0.000118     8485.
##  6 num_hrefs                0.577           1.73
##  7 num_self_hrefs           0.706           1.42
##  8 num_imgs                 0.604           1.65
##  9 num_videos               0.797           1.25
## 10 average_token_length     0.0793         12.6
## # ... with 49 more rows
```

VARIABLE SELECTION METHODS

```
#Fwd stepwise:
fwd_fit <- regsubsets(popular_article ~. ,
             data = binomial_train,
             nvmax = 12,
             method = "forward",
             )
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 3 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
summary(fwd_fit)
```

```
## Subset selection object
## Call: regsubsets.formula(popular_article ~ ., data = binomial_train,
##     nvmax = 12, method = "forward", )
## 58 Variables  (and intercept)
##                                Forced in Forced out
## n_tokens_title                     FALSE      FALSE
## n_tokens_content                   FALSE      FALSE
## n_unique_tokens                    FALSE      FALSE
## n_non_stop_words                   FALSE      FALSE
## n_non_stop_unique_tokens           FALSE      FALSE
## num_hrefs                          FALSE      FALSE
## num_self_hrefs                     FALSE      FALSE
## num_imgs                           FALSE      FALSE
## num_videos                         FALSE      FALSE
## average_token_length               FALSE      FALSE
## num_keywords                       FALSE      FALSE
## data_channel_is_lifestyle          FALSE      FALSE
## data_channel_is_entertainment      FALSE      FALSE
## data_channel_is_bus                FALSE      FALSE
## data_channel_is_socmed             FALSE      FALSE
## data_channel_is_tech               FALSE      FALSE
## data_channel_is_world              FALSE      FALSE
## kw_min_min                         FALSE      FALSE
## kw_max_min                         FALSE      FALSE
## kw_avg_min                         FALSE      FALSE
## kw_min_max                         FALSE      FALSE
## kw_max_max                         FALSE      FALSE
```

```
## kw_avg_max                          FALSE      FALSE
## kw_min_avg                          FALSE      FALSE
## kw_max_avg                          FALSE      FALSE
## kw_avg_avg                          FALSE      FALSE
## self_reference_min_shares           FALSE      FALSE
## self_reference_max_shares           FALSE      FALSE
## self_reference_avg_sharess          FALSE      FALSE
## weekday_is_monday                   FALSE      FALSE
## weekday_is_tuesday                  FALSE      FALSE
## weekday_is_wednesday                FALSE      FALSE
## weekday_is_thursday                 FALSE      FALSE
## weekday_is_friday                   FALSE      FALSE
## weekday_is_saturday                 FALSE      FALSE
## LDA_00                              FALSE      FALSE
## LDA_01                              FALSE      FALSE
## LDA_02                              FALSE      FALSE
## LDA_03                              FALSE      FALSE
## global_subjectivity                 FALSE      FALSE
## global_sentiment_polarity           FALSE      FALSE
## global_rate_positive_words          FALSE      FALSE
## global_rate_negative_words          FALSE      FALSE
## rate_positive_words                 FALSE      FALSE
## rate_negative_words                 FALSE      FALSE
## avg_positive_polarity               FALSE      FALSE
## min_positive_polarity               FALSE      FALSE
## max_positive_polarity               FALSE      FALSE
## avg_negative_polarity               FALSE      FALSE
## min_negative_polarity               FALSE      FALSE
## max_negative_polarity               FALSE      FALSE
## title_subjectivity                  FALSE      FALSE
## title_sentiment_polarity            FALSE      FALSE
## abs_title_subjectivity              FALSE      FALSE
## abs_title_sentiment_polarity        FALSE      FALSE
## weekday_is_sunday                   FALSE      FALSE
## is_weekend                          FALSE      FALSE
## LDA_04                              FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: forward
##           n_tokens_title n_tokens_content n_unique_tokens n_non_stop_words
## 1  ( 1 )  " "            " "              " "             " "
## 2  ( 1 )  " "            " "              " "             " "
## 3  ( 1 )  " "            " "              " "             " "
## 4  ( 1 )  " "            " "              " "             " "
## 5  ( 1 )  " "            " "              " "             " "
## 6  ( 1 )  " "            " "              " "             " "
## 7  ( 1 )  " "            " "              " "             " "
## 8  ( 1 )  " "            " "              " "             " "
## 9  ( 1 )  " "            "*"              " "             " "
## 10  ( 1 ) " "            "*"              " "             " "
## 11  ( 1 ) " "            "*"              " "             " "
## 12  ( 1 ) " "            "*"              " "             " "
## 13  ( 1 ) " "            "*"              " "             " "
##           n_non_stop_unique_tokens num_hrefs num_self_hrefs num_imgs
## 1  ( 1 )  " "                      " "       " "            " "
```

```
## 2  ( 1 )  " "                      " "       " "            " "
## 3  ( 1 )  " "                      " "       " "            " "
## 4  ( 1 )  " "                      " "       " "            " "
## 5  ( 1 )  " "                      " "       " "            " "
## 6  ( 1 )  " "                      " "       " "            " "
## 7  ( 1 )  " "                      " "       " "            " "
## 8  ( 1 )  " "                      " "       " "            " "
## 9  ( 1 )  " "                      " "       " "            " "
## 10  ( 1 ) " "                      " "       " "            " "
## 11  ( 1 ) " "                      " "       " "            " "
## 12  ( 1 ) " "                      " "       " "            " "
## 13  ( 1 ) " "                      " "       " "            " "
##           num_videos average_token_length num_keywords
## 1  ( 1 )  " "        " "                  " "
## 2  ( 1 )  " "        " "                  " "
## 3  ( 1 )  " "        " "                  " "
## 4  ( 1 )  " "        " "                  " "
## 5  ( 1 )  " "        " "                  " "
## 6  ( 1 )  " "        " "                  " "
## 7  ( 1 )  " "        " "                  " "
## 8  ( 1 )  " "        " "                  " "
## 9  ( 1 )  " "        " "                  " "
## 10  ( 1 ) " "        " "                  " "
## 11  ( 1 ) " "        " "                  "*"
## 12  ( 1 ) " "        " "                  "*"
## 13  ( 1 ) " "        " "                  "*"
##           data_channel_is_lifestyle data_channel_is_entertainment
## 1  ( 1 )  " "                       " "
## 2  ( 1 )  " "                       " "
## 3  ( 1 )  " "                       " "
## 4  ( 1 )  " "                       "*"
## 5  ( 1 )  " "                       "*"
## 6  ( 1 )  " "                       "*"
## 7  ( 1 )  " "                       "*"
## 8  ( 1 )  " "                       "*"
## 9  ( 1 )  " "                       "*"
## 10  ( 1 ) " "                       "*"
## 11  ( 1 ) " "                       "*"
## 12  ( 1 ) " "                       "*"
## 13  ( 1 ) " "                       "*"
##           data_channel_is_bus data_channel_is_socmed data_channel_is_tech
## 1  ( 1 )  " "                 " "                    " "
## 2  ( 1 )  " "                 " "                    " "
## 3  ( 1 )  " "                 " "                    " "
## 4  ( 1 )  " "                 " "                    " "
## 5  ( 1 )  " "                 "*"                    " "
## 6  ( 1 )  " "                 "*"                    "*"
## 7  ( 1 )  " "                 "*"                    "*"
## 8  ( 1 )  " "                 "*"                    "*"
## 9  ( 1 )  " "                 "*"                    "*"
## 10  ( 1 ) " "                 "*"                    "*"
## 11  ( 1 ) " "                 "*"                    "*"
## 12  ( 1 ) " "                 "*"                    "*"
## 13  ( 1 ) " "                 "*"                    "*"
```

```
##           data_channel_is_world kw_min_min kw_max_min kw_avg_min
## 1  ( 1 )  " "                   " "        " "        " "
## 2  ( 1 )  " "                   " "        " "        " "
## 3  ( 1 )  " "                   " "        " "        " "
## 4  ( 1 )  " "                   " "        " "        " "
## 5  ( 1 )  " "                   " "        " "        " "
## 6  ( 1 )  " "                   " "        " "        " "
## 7  ( 1 )  " "                   " "        " "        " "
## 8  ( 1 )  " "                   " "        " "        " "
## 9  ( 1 )  " "                   " "        " "        " "
## 10  ( 1 ) " "                   "*"        " "        " "
## 11  ( 1 ) " "                   "*"        " "        " "
## 12  ( 1 ) " "                   "*"        " "        " "
## 13  ( 1 ) " "                   "*"        " "        " "
##           kw_min_max kw_max_max kw_avg_max kw_min_avg kw_max_avg
## 1  ( 1 )  " "        " "        " "        " "        " "
## 2  ( 1 )  " "        " "        " "        " "        " "
## 3  ( 1 )  " "        " "        " "        " "        " "
## 4  ( 1 )  " "        " "        " "        " "        " "
## 5  ( 1 )  " "        " "        " "        " "        " "
## 6  ( 1 )  " "        " "        " "        " "        " "
## 7  ( 1 )  " "        " "        " "        " "        "*"
## 8  ( 1 )  " "        " "        "*"        " "        "*"
## 9  ( 1 )  " "        " "        "*"        " "        "*"
## 10  ( 1 ) " "        " "        "*"        " "        "*"
## 11  ( 1 ) " "        " "        "*"        " "        "*"
## 12  ( 1 ) " "        " "        "*"        " "        "*"
## 13  ( 1 ) " "        " "        "*"        " "        "*"
##           kw_avg_avg self_reference_min_shares self_reference_max_shares
## 1  ( 1 )  "*"        " "                       " "
## 2  ( 1 )  "*"        " "                       " "
## 3  ( 1 )  "*"        " "                       " "
## 4  ( 1 )  "*"        " "                       " "
## 5  ( 1 )  "*"        " "                       " "
## 6  ( 1 )  "*"        " "                       " "
## 7  ( 1 )  "*"        " "                       " "
## 8  ( 1 )  "*"        " "                       " "
## 9  ( 1 )  "*"        " "                       " "
## 10  ( 1 ) "*"        " "                       " "
## 11  ( 1 ) "*"        " "                       " "
## 12  ( 1 ) "*"        " "                       " "
## 13  ( 1 ) "*"        " "                       " "
##           self_reference_avg_sharess weekday_is_monday weekday_is_tuesday
## 1  ( 1 )  " "                        " "               " "
## 2  ( 1 )  " "                        " "               " "
## 3  ( 1 )  " "                        " "               " "
## 4  ( 1 )  " "                        " "               " "
## 5  ( 1 )  " "                        " "               " "
## 6  ( 1 )  " "                        " "               " "
## 7  ( 1 )  " "                        " "               " "
## 8  ( 1 )  " "                        " "               " "
## 9  ( 1 )  " "                        " "               " "
## 10  ( 1 ) " "                        " "               " "
## 11  ( 1 ) " "                        " "               " "
```

```
## 12  ( 1 ) " "                         " "                " "
## 13  ( 1 ) "*"                         " "                " "
##           weekday_is_wednesday weekday_is_thursday weekday_is_friday
## 1  ( 1 )  " "                  " "                 " "
## 2  ( 1 )  " "                  " "                 " "
## 3  ( 1 )  " "                  " "                 " "
## 4  ( 1 )  " "                  " "                 " "
## 5  ( 1 )  " "                  " "                 " "
## 6  ( 1 )  " "                  " "                 " "
## 7  ( 1 )  " "                  " "                 " "
## 8  ( 1 )  " "                  " "                 " "
## 9  ( 1 )  " "                  " "                 " "
## 10  ( 1 ) " "                  " "                 " "
## 11  ( 1 ) " "                  " "                 " "
## 12  ( 1 ) " "                  " "                 " "
## 13  ( 1 ) " "                  " "                 " "
##           weekday_is_saturday weekday_is_sunday is_weekend LDA_00 LDA_01
## 1  ( 1 )  " "                 " "               " "        " "    " "
## 2  ( 1 )  " "                 " "               "*"        " "    " "
## 3  ( 1 )  " "                 " "               "*"        " "    " "
## 4  ( 1 )  " "                 " "               "*"        " "    " "
## 5  ( 1 )  " "                 " "               "*"        " "    " "
## 6  ( 1 )  " "                 " "               "*"        " "    " "
## 7  ( 1 )  " "                 " "               "*"        " "    " "
## 8  ( 1 )  " "                 " "               "*"        " "    " "
## 9  ( 1 )  " "                 " "               "*"        " "    " "
## 10  ( 1 ) " "                 " "               "*"        " "    " "
## 11  ( 1 ) " "                 " "               "*"        " "    " "
## 12  ( 1 ) " "                 " "               "*"        "*"    " "
## 13  ( 1 ) " "                 " "               "*"        "*"    " "
##           LDA_02 LDA_03 LDA_04 global_subjectivity
## 1  ( 1 )  " "    " "    " "    " "
## 2  ( 1 )  " "    " "    " "    " "
## 3  ( 1 )  "*"    " "    " "    " "
## 4  ( 1 )  "*"    " "    " "    " "
## 5  ( 1 )  "*"    " "    " "    " "
## 6  ( 1 )  "*"    " "    " "    " "
## 7  ( 1 )  "*"    " "    " "    " "
## 8  ( 1 )  "*"    " "    " "    " "
## 9  ( 1 )  "*"    " "    " "    " "
## 10  ( 1 ) "*"    " "    " "    " "
## 11  ( 1 ) "*"    " "    " "    " "
## 12  ( 1 ) "*"    " "    " "    " "
## 13  ( 1 ) "*"    " "    " "    " "
##           global_sentiment_polarity global_rate_positive_words
## 1  ( 1 )  " "                       " "
## 2  ( 1 )  " "                       " "
## 3  ( 1 )  " "                       " "
## 4  ( 1 )  " "                       " "
## 5  ( 1 )  " "                       " "
## 6  ( 1 )  " "                       " "
## 7  ( 1 )  " "                       " "
## 8  ( 1 )  " "                       " "
## 9  ( 1 )  " "                       " "
```

```
## 10  ( 1 ) " "                        " "
## 11  ( 1 ) " "                        " "
## 12  ( 1 ) " "                        " "
## 13  ( 1 ) " "                        " "
##           global_rate_negative_words rate_positive_words
## 1  ( 1 )  " "                         " "
## 2  ( 1 )  " "                         " "
## 3  ( 1 )  " "                         " "
## 4  ( 1 )  " "                         " "
## 5  ( 1 )  " "                         " "
## 6  ( 1 )  " "                         " "
## 7  ( 1 )  " "                         " "
## 8  ( 1 )  " "                         " "
## 9  ( 1 )  " "                         " "
## 10  ( 1 ) " "                         " "
## 11  ( 1 ) " "                         " "
## 12  ( 1 ) " "                         " "
## 13  ( 1 ) " "                         " "
##           rate_negative_words avg_positive_polarity min_positive_polarity
## 1  ( 1 )  " "                 " "                   " "
## 2  ( 1 )  " "                 " "                   " "
## 3  ( 1 )  " "                 " "                   " "
## 4  ( 1 )  " "                 " "                   " "
## 5  ( 1 )  " "                 " "                   " "
## 6  ( 1 )  " "                 " "                   " "
## 7  ( 1 )  " "                 " "                   " "
## 8  ( 1 )  " "                 " "                   " "
## 9  ( 1 )  " "                 " "                   " "
## 10  ( 1 ) " "                 " "                   " "
## 11  ( 1 ) " "                 " "                   " "
## 12  ( 1 ) " "                 " "                   " "
## 13  ( 1 ) " "                 " "                   " "
##           max_positive_polarity avg_negative_polarity
## 1  ( 1 )  " "                   " "
## 2  ( 1 )  " "                   " "
## 3  ( 1 )  " "                   " "
## 4  ( 1 )  " "                   " "
## 5  ( 1 )  " "                   " "
## 6  ( 1 )  " "                   " "
## 7  ( 1 )  " "                   " "
## 8  ( 1 )  " "                   " "
## 9  ( 1 )  " "                   " "
## 10  ( 1 ) " "                   " "
## 11  ( 1 ) " "                   " "
## 12  ( 1 ) " "                   " "
## 13  ( 1 ) " "                   " "
##           min_negative_polarity max_negative_polarity title_subjectivity
## 1  ( 1 )  " "                   " "                   " "
## 2  ( 1 )  " "                   " "                   " "
## 3  ( 1 )  " "                   " "                   " "
## 4  ( 1 )  " "                   " "                   " "
## 5  ( 1 )  " "                   " "                   " "
## 6  ( 1 )  " "                   " "                   " "
## 7  ( 1 )  " "                   " "                   " "
```

```
## 8  ( 1 )  " "                   " "                   " "
## 9  ( 1 )  " "                   " "                   " "
## 10  ( 1 ) " "                   " "                   " "
## 11  ( 1 ) " "                   " "                   " "
## 12  ( 1 ) " "                   " "                   " "
## 13  ( 1 ) " "                   " "                   " "
##           title_sentiment_polarity abs_title_subjectivity
## 1  ( 1 )  " "                       " "
## 2  ( 1 )  " "                       " "
## 3  ( 1 )  " "                       " "
## 4  ( 1 )  " "                       " "
## 5  ( 1 )  " "                       " "
## 6  ( 1 )  " "                       " "
## 7  ( 1 )  " "                       " "
## 8  ( 1 )  " "                       " "
## 9  ( 1 )  " "                       " "
## 10  ( 1 ) " "                       " "
## 11  ( 1 ) " "                       " "
## 12  ( 1 ) " "                       " "
## 13  ( 1 ) " "                       " "
##           abs_title_sentiment_polarity
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
## 11  ( 1 ) " "
## 12  ( 1 ) " "
## 13  ( 1 ) " "
```

```
#Bkwd stepwise:
bkwd_fit <- regsubsets(popular_article ~. ,
                       data = binomial_train,
                       nvmax = 12,
                       method = "backward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 3 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
summary(bkwd_fit)
```

```
## Subset selection object
## Call: regsubsets.formula(popular_article ~ ., data = binomial_train,
##     nvmax = 12, method = "backward")
## 58 Variables  (and intercept)
##                               Forced in Forced out
## n_tokens_title                    FALSE      FALSE
## n_tokens_content                  FALSE      FALSE
## n_unique_tokens                   FALSE      FALSE
```

```
## n_non_stop_words                 FALSE      FALSE
## n_non_stop_unique_tokens         FALSE      FALSE
## num_hrefs                        FALSE      FALSE
## num_self_hrefs                   FALSE      FALSE
## num_imgs                         FALSE      FALSE
## num_videos                       FALSE      FALSE
## average_token_length             FALSE      FALSE
## num_keywords                     FALSE      FALSE
## data_channel_is_lifestyle        FALSE      FALSE
## data_channel_is_entertainment    FALSE      FALSE
## data_channel_is_bus              FALSE      FALSE
## data_channel_is_socmed           FALSE      FALSE
## data_channel_is_tech             FALSE      FALSE
## data_channel_is_world            FALSE      FALSE
## kw_min_min                       FALSE      FALSE
## kw_max_min                       FALSE      FALSE
## kw_avg_min                       FALSE      FALSE
## kw_min_max                       FALSE      FALSE
## kw_max_max                       FALSE      FALSE
## kw_avg_max                       FALSE      FALSE
## kw_min_avg                       FALSE      FALSE
## kw_max_avg                       FALSE      FALSE
## kw_avg_avg                       FALSE      FALSE
## self_reference_min_shares        FALSE      FALSE
## self_reference_max_shares        FALSE      FALSE
## self_reference_avg_sharess       FALSE      FALSE
## weekday_is_monday                FALSE      FALSE
## weekday_is_tuesday               FALSE      FALSE
## weekday_is_wednesday             FALSE      FALSE
## weekday_is_thursday              FALSE      FALSE
## weekday_is_friday                FALSE      FALSE
## weekday_is_saturday              FALSE      FALSE
## LDA_00                           FALSE      FALSE
## LDA_01                           FALSE      FALSE
## LDA_02                           FALSE      FALSE
## LDA_03                           FALSE      FALSE
## global_subjectivity              FALSE      FALSE
## global_sentiment_polarity        FALSE      FALSE
## global_rate_positive_words       FALSE      FALSE
## global_rate_negative_words       FALSE      FALSE
## rate_positive_words              FALSE      FALSE
## rate_negative_words              FALSE      FALSE
## avg_positive_polarity            FALSE      FALSE
## min_positive_polarity            FALSE      FALSE
## max_positive_polarity            FALSE      FALSE
## avg_negative_polarity            FALSE      FALSE
## min_negative_polarity            FALSE      FALSE
## max_negative_polarity            FALSE      FALSE
## title_subjectivity               FALSE      FALSE
## title_sentiment_polarity         FALSE      FALSE
## abs_title_subjectivity           FALSE      FALSE
## abs_title_sentiment_polarity     FALSE      FALSE
## weekday_is_sunday                FALSE      FALSE
## is_weekend                       FALSE      FALSE
```

```
## LDA_04                              FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: backward
##           n_tokens_title n_tokens_content n_unique_tokens n_non_stop_words
## 1  ( 1 )  " "            " "              " "             " "
## 2  ( 1 )  " "            " "              " "             " "
## 3  ( 1 )  " "            " "              " "             " "
## 4  ( 1 )  " "            " "              " "             " "
## 5  ( 1 )  " "            " "              " "             " "
## 6  ( 1 )  " "            " "              " "             " "
## 7  ( 1 )  " "            " "              " "             " "
## 8  ( 1 )  " "            " "              " "             " "
## 9  ( 1 )  " "            " "              " "             " "
## 10  ( 1 ) " "            " "              " "             " "
## 11  ( 1 ) " "            " "              " "             " "
## 12  ( 1 ) " "            " "              " "             " "
## 13  ( 1 ) " "            " "              " "             " "
##           n_non_stop_unique_tokens num_hrefs num_self_hrefs num_imgs
## 1  ( 1 )  " "                      " "       " "            " "
## 2  ( 1 )  " "                      " "       " "            " "
## 3  ( 1 )  " "                      " "       " "            " "
## 4  ( 1 )  " "                      " "       " "            " "
## 5  ( 1 )  " "                      " "       " "            " "
## 6  ( 1 )  " "                      " "       " "            " "
## 7  ( 1 )  " "                      " "       " "            " "
## 8  ( 1 )  " "                      " "       " "            " "
## 9  ( 1 )  " "                      " "       " "            " "
## 10  ( 1 ) " "                      " "       " "            " "
## 11  ( 1 ) " "                      " "       " "            " "
## 12  ( 1 ) " "                      " "       " "            " "
## 13  ( 1 ) "*"                      " "       " "            " "
##           num_videos average_token_length num_keywords
## 1  ( 1 )  " "        " "                  " "
## 2  ( 1 )  " "        " "                  " "
## 3  ( 1 )  " "        " "                  " "
## 4  ( 1 )  " "        " "                  " "
## 5  ( 1 )  " "        " "                  " "
## 6  ( 1 )  " "        " "                  " "
## 7  ( 1 )  " "        " "                  "*"
## 8  ( 1 )  " "        " "                  "*"
## 9  ( 1 )  " "        " "                  "*"
## 10  ( 1 ) " "        " "                  "*"
## 11  ( 1 ) " "        " "                  "*"
## 12  ( 1 ) " "        " "                  "*"
## 13  ( 1 ) " "        " "                  "*"
##           data_channel_is_lifestyle data_channel_is_entertainment
## 1  ( 1 )  " "                       " "
## 2  ( 1 )  " "                       " "
## 3  ( 1 )  " "                       " "
## 4  ( 1 )  " "                       " "
## 5  ( 1 )  " "                       " "
## 6  ( 1 )  " "                       " "
## 7  ( 1 )  " "                       " "
## 8  ( 1 )  " "                       " "
```

```
## 9  ( 1 )  " "                        " "
## 10  ( 1 ) " "                        " "
## 11  ( 1 ) " "                        " "
## 12  ( 1 ) " "                        " "
## 13  ( 1 ) " "                        " "
##           data_channel_is_bus data_channel_is_socmed data_channel_is_tech
## 1  ( 1 )  " "                 " "                    " "
## 2  ( 1 )  " "                 " "                    "*"
## 3  ( 1 )  " "                 "*"                    "*"
## 4  ( 1 )  " "                 "*"                    "*"
## 5  ( 1 )  " "                 "*"                    "*"
## 6  ( 1 )  " "                 "*"                    "*"
## 7  ( 1 )  " "                 "*"                    "*"
## 8  ( 1 )  " "                 "*"                    "*"
## 9  ( 1 )  " "                 "*"                    "*"
## 10  ( 1 ) " "                 "*"                    "*"
## 11  ( 1 ) " "                 "*"                    "*"
## 12  ( 1 ) " "                 "*"                    "*"
## 13  ( 1 ) " "                 "*"                    "*"
##           data_channel_is_world kw_min_min kw_max_min kw_avg_min
## 1  ( 1 )  " "                   " "        " "        " "
## 2  ( 1 )  " "                   " "        " "        " "
## 3  ( 1 )  " "                   " "        " "        " "
## 4  ( 1 )  " "                   " "        " "        " "
## 5  ( 1 )  " "                   "*"        " "        " "
## 6  ( 1 )  " "                   "*"        " "        " "
## 7  ( 1 )  " "                   "*"        " "        " "
## 8  ( 1 )  " "                   "*"        " "        " "
## 9  ( 1 )  " "                   "*"        " "        " "
## 10  ( 1 ) " "                   "*"        " "        " "
## 11  ( 1 ) " "                   "*"        " "        " "
## 12  ( 1 ) " "                   "*"        " "        " "
## 13  ( 1 ) " "                   "*"        " "        " "
##           kw_min_max kw_max_max kw_avg_max kw_min_avg kw_max_avg
## 1  ( 1 )  " "        " "        " "        " "        " "
## 2  ( 1 )  " "        " "        " "        " "        " "
## 3  ( 1 )  " "        " "        " "        " "        " "
## 4  ( 1 )  " "        " "        " "        " "        "*"
## 5  ( 1 )  " "        " "        " "        " "        "*"
## 6  ( 1 )  " "        " "        " "        " "        "*"
## 7  ( 1 )  " "        " "        " "        " "        "*"
## 8  ( 1 )  " "        " "        " "        " "        "*"
## 9  ( 1 )  " "        " "        " "        " "        "*"
## 10  ( 1 ) " "        " "        " "        " "        "*"
## 11  ( 1 ) " "        " "        " "        " "        "*"
## 12  ( 1 ) " "        " "        " "        " "        "*"
## 13  ( 1 ) " "        " "        " "        " "        "*"
##           kw_avg_avg self_reference_min_shares self_reference_max_shares
## 1  ( 1 )  "*"        " "                       " "
## 2  ( 1 )  "*"        " "                       " "
## 3  ( 1 )  "*"        " "                       " "
## 4  ( 1 )  "*"        " "                       " "
## 5  ( 1 )  "*"        " "                       " "
## 6  ( 1 )  "*"        " "                       " "
```

```
## 7  ( 1 )  "*"        " "                      " "
## 8  ( 1 )  "*"        " "                      " "
## 9  ( 1 )  "*"        " "                      " "
## 10  ( 1 ) "*"        " "                      " "
## 11  ( 1 ) "*"        " "                      " "
## 12  ( 1 ) "*"        " "                      " "
## 13  ( 1 ) "*"        " "                      " "
##           self_reference_avg_sharess weekday_is_monday weekday_is_tuesday
## 1  ( 1 )  " "                        " "               " "
## 2  ( 1 )  " "                        " "               " "
## 3  ( 1 )  " "                        " "               " "
## 4  ( 1 )  " "                        " "               " "
## 5  ( 1 )  " "                        " "               " "
## 6  ( 1 )  " "                        " "               " "
## 7  ( 1 )  " "                        " "               " "
## 8  ( 1 )  " "                        " "               "*"
## 9  ( 1 )  " "                        " "               "*"
## 10  ( 1 ) " "                        " "               "*"
## 11  ( 1 ) " "                        "*"               "*"
## 12  ( 1 ) " "                        "*"               "*"
## 13  ( 1 ) " "                        "*"               "*"
##           weekday_is_wednesday weekday_is_thursday weekday_is_friday
## 1  ( 1 )  " "                  " "                 " "
## 2  ( 1 )  " "                  " "                 " "
## 3  ( 1 )  " "                  " "                 " "
## 4  ( 1 )  " "                  " "                 " "
## 5  ( 1 )  " "                  " "                 " "
## 6  ( 1 )  " "                  " "                 " "
## 7  ( 1 )  " "                  " "                 " "
## 8  ( 1 )  " "                  " "                 " "
## 9  ( 1 )  "*"                  " "                 " "
## 10  ( 1 ) "*"                  "*"                 " "
## 11  ( 1 ) "*"                  "*"                 " "
## 12  ( 1 ) "*"                  "*"                 "*"
## 13  ( 1 ) "*"                  "*"                 "*"
##           weekday_is_saturday weekday_is_sunday is_weekend LDA_00 LDA_01
## 1  ( 1 )  " "                 " "               " "        " "    " "
## 2  ( 1 )  " "                 " "               " "        " "    " "
## 3  ( 1 )  " "                 " "               " "        " "    " "
## 4  ( 1 )  " "                 " "               " "        " "    " "
## 5  ( 1 )  " "                 " "               " "        " "    " "
## 6  ( 1 )  " "                 " "               " "        "*"    " "
## 7  ( 1 )  " "                 " "               " "        "*"    " "
## 8  ( 1 )  " "                 " "               " "        "*"    " "
## 9  ( 1 )  " "                 " "               " "        "*"    " "
## 10  ( 1 ) " "                 " "               " "        "*"    " "
## 11  ( 1 ) " "                 " "               " "        "*"    " "
## 12  ( 1 ) " "                 " "               " "        "*"    " "
## 13  ( 1 ) " "                 " "               " "        "*"    " "
##           LDA_02 LDA_03 LDA_04 global_subjectivity
## 1  ( 1 )  " "    " "    " "    " "
## 2  ( 1 )  " "    " "    " "    " "
## 3  ( 1 )  " "    " "    " "    " "
## 4  ( 1 )  " "    " "    " "    " "
```

```
## 5  ( 1 )  " "    " "    " "    " "
## 6  ( 1 )  " "    " "    " "    " "
## 7  ( 1 )  " "    " "    " "    " "
## 8  ( 1 )  " "    " "    " "    " "
## 9  ( 1 )  " "    " "    " "    " "
## 10  ( 1 ) " "    " "    " "    " "
## 11  ( 1 ) " "    " "    " "    " "
## 12  ( 1 ) " "    " "    " "    " "
## 13  ( 1 ) " "    " "    " "    " "
##           global_sentiment_polarity global_rate_positive_words
## 1  ( 1 )  " "                       " "
## 2  ( 1 )  " "                       " "
## 3  ( 1 )  " "                       " "
## 4  ( 1 )  " "                       " "
## 5  ( 1 )  " "                       " "
## 6  ( 1 )  " "                       " "
## 7  ( 1 )  " "                       " "
## 8  ( 1 )  " "                       " "
## 9  ( 1 )  " "                       " "
## 10  ( 1 ) " "                       " "
## 11  ( 1 ) " "                       " "
## 12  ( 1 ) " "                       " "
## 13  ( 1 ) " "                       " "
##           global_rate_negative_words rate_positive_words
## 1  ( 1 )  " "                        " "
## 2  ( 1 )  " "                        " "
## 3  ( 1 )  " "                        " "
## 4  ( 1 )  " "                        " "
## 5  ( 1 )  " "                        " "
## 6  ( 1 )  " "                        " "
## 7  ( 1 )  " "                        " "
## 8  ( 1 )  " "                        " "
## 9  ( 1 )  " "                        " "
## 10  ( 1 ) " "                        " "
## 11  ( 1 ) " "                        " "
## 12  ( 1 ) " "                        " "
## 13  ( 1 ) " "                        " "
##           rate_negative_words avg_positive_polarity min_positive_polarity
## 1  ( 1 )  " "                 " "                   " "
## 2  ( 1 )  " "                 " "                   " "
## 3  ( 1 )  " "                 " "                   " "
## 4  ( 1 )  " "                 " "                   " "
## 5  ( 1 )  " "                 " "                   " "
## 6  ( 1 )  " "                 " "                   " "
## 7  ( 1 )  " "                 " "                   " "
## 8  ( 1 )  " "                 " "                   " "
## 9  ( 1 )  " "                 " "                   " "
## 10  ( 1 ) " "                 " "                   " "
## 11  ( 1 ) " "                 " "                   " "
## 12  ( 1 ) " "                 " "                   " "
## 13  ( 1 ) " "                 " "                   " "
##           max_positive_polarity avg_negative_polarity
## 1  ( 1 )  " "                   " "
## 2  ( 1 )  " "                   " "
```

```
## 3  ( 1 )  " "                   " "
## 4  ( 1 )  " "                   " "
## 5  ( 1 )  " "                   " "
## 6  ( 1 )  " "                   " "
## 7  ( 1 )  " "                   " "
## 8  ( 1 )  " "                   " "
## 9  ( 1 )  " "                   " "
## 10  ( 1 ) " "                   " "
## 11  ( 1 ) " "                   " "
## 12  ( 1 ) " "                   " "
## 13  ( 1 ) " "                   " "
##           min_negative_polarity max_negative_polarity title_subjectivity
## 1  ( 1 )  " "                   " "                   " "
## 2  ( 1 )  " "                   " "                   " "
## 3  ( 1 )  " "                   " "                   " "
## 4  ( 1 )  " "                   " "                   " "
## 5  ( 1 )  " "                   " "                   " "
## 6  ( 1 )  " "                   " "                   " "
## 7  ( 1 )  " "                   " "                   " "
## 8  ( 1 )  " "                   " "                   " "
## 9  ( 1 )  " "                   " "                   " "
## 10  ( 1 ) " "                   " "                   " "
## 11  ( 1 ) " "                   " "                   " "
## 12  ( 1 ) " "                   " "                   " "
## 13  ( 1 ) " "                   " "                   " "
##           title_sentiment_polarity abs_title_subjectivity
## 1  ( 1 )  " "                      " "
## 2  ( 1 )  " "                      " "
## 3  ( 1 )  " "                      " "
## 4  ( 1 )  " "                      " "
## 5  ( 1 )  " "                      " "
## 6  ( 1 )  " "                      " "
## 7  ( 1 )  " "                      " "
## 8  ( 1 )  " "                      " "
## 9  ( 1 )  " "                      " "
## 10  ( 1 ) " "                      " "
## 11  ( 1 ) " "                      " "
## 12  ( 1 ) " "                      " "
## 13  ( 1 ) " "                      " "
##           abs_title_sentiment_polarity
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
## 11  ( 1 ) " "
## 12  ( 1 ) " "
## 13  ( 1 ) " "
```

Ridge and Lasso Mods

```
#Ridge-regression:
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loaded glmnet 2.0-18
```

```
library(glmnetUtils)
```

```
##
## Attaching package: 'glmnetUtils'
```

```
## The following objects are masked from 'package:glmnet':
##
##     cv.glmnet, glmnet
```

```
ridge_mod <- cv.glmnet(popular_article ~. ,
                       data = binomial_train,
                       alpha = 0)
coef(ridge_mod)
```

```
## 59 x 1 sparse Matrix of class "dgCMatrix"
##                                        1
## (Intercept)                 2.526519e-01
## n_tokens_title             -2.222498e-05
## n_tokens_content            1.869792e-05
## n_unique_tokens            -3.521947e-02
## n_non_stop_words           -1.236763e-02
## n_non_stop_unique_tokens   -1.052859e-01
## num_hrefs                   1.895315e-03
## num_self_hrefs             -5.002273e-03
## num_imgs                    8.783236e-04
## num_videos                  4.663704e-04
## average_token_length       -9.547174e-03
## num_keywords                1.151283e-02
## data_channel_is_lifestyle  -2.875689e-02
## data_channel_is_entertainment -9.962570e-02
## data_channel_is_bus        -7.812597e-02
## data_channel_is_socmed      1.430882e-01
## data_channel_is_tech        7.462413e-02
## data_channel_is_world      -3.825646e-02
```

```
## kw_min_min                    3.517829e-04
## kw_max_min                    6.190136e-07
## kw_avg_min                   -9.848595e-06
## kw_min_max                   -1.608605e-07
## kw_max_max                   -4.623504e-08
## kw_avg_max                   -1.645471e-08
## kw_min_avg                    8.952549e-06
## kw_max_avg                   -8.311775e-06
## kw_avg_avg                    7.750642e-05
## self_reference_min_shares     3.336517e-07
## self_reference_max_shares     8.779531e-08
## self_reference_avg_sharess    3.097380e-07
## weekday_is_monday            -2.275587e-03
## weekday_is_tuesday           -3.134507e-02
## weekday_is_wednesday         -2.992062e-02
## weekday_is_thursday          -1.209904e-02
## weekday_is_friday             1.877455e-02
## weekday_is_saturday           1.021957e-01
## weekday_is_sunday             5.136462e-02
## is_weekend                    8.107055e-02
## LDA_00                        1.517568e-01
## LDA_01                       -5.574073e-02
## LDA_02                       -1.175538e-01
## LDA_03                       -2.326996e-02
## LDA_04                        4.272525e-02
## global_subjectivity           2.479573e-01
## global_sentiment_polarity     4.526512e-03
## global_rate_positive_words   -3.025441e-01
## global_rate_negative_words    7.341441e-01
## rate_positive_words           2.831946e-02
## rate_negative_words          -3.630336e-02
## avg_positive_polarity        -9.371393e-02
## min_positive_polarity        -3.038494e-02
## max_positive_polarity         5.253668e-03
## avg_negative_polarity        -5.279181e-03
## min_negative_polarity         4.802346e-03
## max_negative_polarity        -5.251446e-03
## title_subjectivity            2.488416e-02
## title_sentiment_polarity      5.684374e-02
## abs_title_subjectivity        5.340125e-02
## abs_title_sentiment_polarity -2.062428e-03
```

```
#Lasso model: More Severe Penalty Term
lasso_mod <- cv.glmnet(popular_article ~. ,
                       data = binomial_train,
                       alpha = 1)
coef(lasso_mod)
```

```
## 59 x 1 sparse Matrix of class "dgCMatrix"
##                                        1
## (Intercept)                1.816568e-01
## n_tokens_title             .
## n_tokens_content           1.101944e-05
## n_unique_tokens           -1.015052e-02
## n_non_stop_words           .
```

```
## n_non_stop_unique_tokens      -1.563260e-01
## num_hrefs                      1.532635e-03
## num_self_hrefs                -2.909661e-03
## num_imgs                       2.789829e-04
## num_videos                     .
## average_token_length           .
## num_keywords                   1.051306e-02
## data_channel_is_lifestyle      .
## data_channel_is_entertainment -8.489859e-02
## data_channel_is_bus           -2.802730e-02
## data_channel_is_socmed         1.686656e-01
## data_channel_is_tech           1.066965e-01
## data_channel_is_world          .
## kw_min_min                     3.770134e-04
## kw_max_min                     .
## kw_avg_min                     .
## kw_min_max                    -9.597610e-08
## kw_max_max                    -3.193796e-08
## kw_avg_max                    -7.130434e-09
## kw_min_avg                     .
## kw_max_avg                    -1.072027e-05
## kw_avg_avg                     9.312038e-05
## self_reference_min_shares      1.431879e-07
## self_reference_max_shares      .
## self_reference_avg_sharess     4.109744e-07
## weekday_is_monday              .
## weekday_is_tuesday            -1.485798e-02
## weekday_is_wednesday          -1.420194e-02
## weekday_is_thursday            .
## weekday_is_friday              1.870615e-02
## weekday_is_saturday            4.437739e-02
## weekday_is_sunday              .
## is_weekend                     1.385266e-01
## LDA_00                         1.167233e-01
## LDA_01                        -2.590141e-02
## LDA_02                        -1.204233e-01
## LDA_03                         .
## LDA_04                         2.942776e-02
## global_subjectivity            1.830557e-01
## global_sentiment_polarity      .
## global_rate_positive_words     .
## global_rate_negative_words     .
## rate_positive_words            .
## rate_negative_words           -9.482154e-04
## avg_positive_polarity         -8.326089e-05
## min_positive_polarity         -3.574673e-02
## max_positive_polarity          .
## avg_negative_polarity          .
## min_negative_polarity          .
## max_negative_polarity          .
## title_subjectivity             5.563975e-03
## title_sentiment_polarity       4.463403e-02
## abs_title_subjectivity         1.345398e-02
## abs_title_sentiment_polarity   .
```

```r
#Coefficient table at lambda.min and lambda.1se for LASSO
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
lasso_coefs <- data.table(
  varnames = rownames(coef(lasso_mod, s = "lambda.min")),
  lasso_lambda_min = round(as.matrix(coef(lasso_mod, s = "lambda.min")), digits = 3), lasso_lambda_1se =
print(lasso_coefs)
```

```
##                         varnames lasso_lambda_min.1 lasso_lambda_1se.1
##  1:                   (Intercept)              0.103              0.182
##  2:                 n_tokens_title              0.001              0.000
##  3:               n_tokens_content              0.000              0.000
##  4:                n_unique_tokens              0.001             -0.010
##  5:               n_non_stop_words              0.000              0.000
##  6:      n_non_stop_unique_tokens             -0.146             -0.156
##  7:                      num_hrefs              0.002              0.002
##  8:                 num_self_hrefs             -0.005             -0.003
##  9:                       num_imgs              0.001              0.000
## 10:                     num_videos              0.001              0.000
## 11:           average_token_length             -0.015              0.000
## 12:                   num_keywords              0.010              0.011
## 13:     data_channel_is_lifestyle             -0.032              0.000
## 14: data_channel_is_entertainment             -0.084             -0.085
## 15:            data_channel_is_bus             -0.076             -0.028
## 16:         data_channel_is_socmed              0.149              0.169
## 17:           data_channel_is_tech              0.094              0.107
## 18:          data_channel_is_world             -0.014              0.000
## 19:                     kw_min_min              0.000              0.000
## 20:                     kw_max_min              0.000              0.000
## 21:                     kw_avg_min              0.000              0.000
## 22:                     kw_min_max              0.000              0.000
## 23:                     kw_max_max              0.000              0.000
## 24:                     kw_avg_max              0.000              0.000
## 25:                     kw_min_avg              0.000              0.000
## 26:                     kw_max_avg              0.000              0.000
## 27:                     kw_avg_avg              0.000              0.000
## 28:     self_reference_min_shares              0.000              0.000
## 29:     self_reference_max_shares              0.000              0.000
## 30:    self_reference_avg_sharess              0.000              0.000
## 31:              weekday_is_monday              0.011              0.000
## 32:             weekday_is_tuesday             -0.020             -0.015
## 33:           weekday_is_wednesday             -0.018             -0.014
## 34:            weekday_is_thursday              0.000              0.000
## 35:              weekday_is_friday              0.032              0.019
```

```
## 36:           weekday_is_saturday              0.051              0.044
## 37:             weekday_is_sunday              0.000              0.000
## 38:                    is_weekend              0.147              0.139
## 39:                        LDA_00              0.232              0.117
## 40:                        LDA_01             -0.013             -0.026
## 41:                        LDA_02             -0.060             -0.120
## 42:                        LDA_03              0.000              0.000
## 43:                        LDA_04              0.095              0.029
## 44:           global_subjectivity              0.259              0.183
## 45:    global_sentiment_polarity              0.006              0.000
## 46:   global_rate_positive_words             -0.488              0.000
## 47:   global_rate_negative_words              0.924              0.000
## 48:           rate_positive_words              0.073              0.000
## 49:           rate_negative_words             -0.007             -0.001
## 50:         avg_positive_polarity             -0.122              0.000
## 51:         min_positive_polarity             -0.026             -0.036
## 52:         max_positive_polarity              0.008              0.000
## 53:         avg_negative_polarity              0.000              0.000
## 54:         min_negative_polarity              0.005              0.000
## 55:         max_negative_polarity             -0.004              0.000
## 56:            title_subjectivity              0.025              0.006
## 57:      title_sentiment_polarity              0.059              0.045
## 58:        abs_title_subjectivity              0.056              0.013
## 59:  abs_title_sentiment_polarity             -0.004              0.000
##                          varnames lasso_lambda_min.1 lasso_lambda_1se.1
```

Elastic Net Model

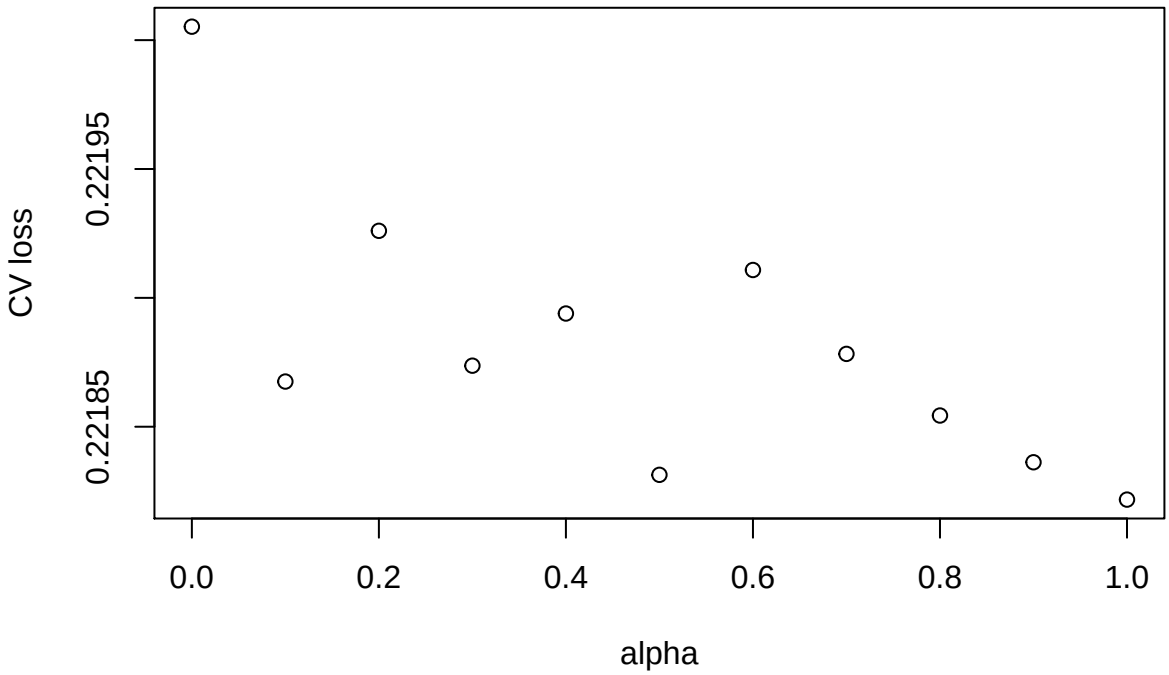
```
alpha_grid <- seq(0,1,len = 11)
alpha_grid
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

```
enet_fit <- cva.glmnet(popular_article ~. ,
                       data = binomial_train,
                       alpha = alpha_grid)
print(enet_fit)
```

```
## Call:
## cva.glmnet.formula(formula = popular_article ~ ., data = binomial_train,
##     alpha = alpha_grid)
##
## Model fitting options:
##     Sparse model matrix: FALSE
##     Use model.frame: FALSE
##     Alpha values: 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
##     Number of crossvalidation folds for lambda: 10
```

```
#plot
minlossplot(enet_fit)
```




```
#matrix of coefficients at each alpha
enet_coefs <- data.frame(
  varname = rownames(coef(enet_fit,alpha = 0)),
  ridge = as.matrix(coef(enet_fit, alpha = 0)) %>% round(3),
  alpha01 = as.matrix(coef(enet_fit, alpha = 0.1)) %>% round(3),
  alpha02 = as.matrix(coef(enet_fit, alpha = 0.2)) %>% round(3),
  alpha03 = as.matrix(coef(enet_fit, alpha = 0.3)) %>% round(3),
  alpha04 = as.matrix(coef(enet_fit, alpha = 0.4)) %>% round(3),
  alpha05 = as.matrix(coef(enet_fit, alpha = 0.5)) %>% round(3),
  alpha06 = as.matrix(coef(enet_fit, alpha = 0.6)) %>% round(3),
  alpha07 = as.matrix(coef(enet_fit, alpha = 0.7)) %>% round(3),
  alpha08 = as.matrix(coef(enet_fit, alpha = 0.8)) %>% round(3),
  alpha09 = as.matrix(coef(enet_fit, alpha = 0.9)) %>% round(3),
  lasso = as.matrix(coef(enet_fit, alpha = 1)) %>% round(3)
) %>% rename(varname = 1, ridge = 2, alpha01 = 3, alpha02 = 4, alpha03 = 5, alpha04 = 6,
             alpha05 = 7, alpha06 = 8, alpha07 = 9, alpha08 = 10, alpha09 = 11, lasso = 12) %>%
  remove_rownames()
head(enet_coefs)
```

```
##                   varname  ridge alpha01 alpha02 alpha03 alpha04 alpha05
## 1             (Intercept)  0.267   0.214   0.205   0.193   0.189   0.183
## 2          n_tokens_title  0.000   0.000   0.000   0.000   0.000   0.000
## 3        n_tokens_content  0.000   0.000   0.000   0.000   0.000   0.000
## 4         n_unique_tokens -0.041  -0.030  -0.028  -0.023  -0.021  -0.017
## 5        n_non_stop_words -0.015   0.000   0.000   0.000   0.000   0.000
## 6 n_non_stop_unique_tokens -0.099  -0.121  -0.130  -0.137  -0.142  -0.146
##   alpha06 alpha07 alpha08 alpha09  lasso
## 1   0.185   0.181   0.179   0.177  0.175
## 2   0.000   0.000   0.000   0.000  0.000
## 3   0.000   0.000   0.000   0.000  0.000
## 4  -0.015  -0.012  -0.009  -0.010 -0.008
## 5   0.000   0.000   0.000   0.000  0.000
## 6  -0.151  -0.153  -0.156  -0.156 -0.158
```

SUMMARY STATS AND PLOTS AFTER VARIABLE SELECTION

```
#FINAL DATASET
#2 Potential Target Variables (shares & popular_article)
#17 Predictor Variables...based off of the minlossplot: we chose the alpha with the lowest CV loss and
#This is what we got:

clean_data <- select(raw_data, c(popular_article,
                                 shares,
                                 num_hrefs,
                                 num_self_hrefs,
                                 num_keywords,
                                 data_channel_is_entertainment,
                                 data_channel_is_bus,
                                 data_channel_is_socmed,
                                 data_channel_is_tech,
                                 weekday_is_tuesday,
                                 weekday_is_wednesday,
                                 weekday_is_friday,
                                 weekday_is_saturday,
                                 is_weekend,
                                 global_subjectivity,
                                 min_positive_polarity,
                                 title_subjectivity,
                                 title_sentiment_polarity,
                                 abs_title_subjectivity))
summary(clean_data)
```

```
##  popular_article      shares         num_hrefs      num_self_hrefs
##  Min.   :0.0000   Min.   :     1   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:0.0000   1st Qu.:   946   1st Qu.:  4.00   1st Qu.:  1.000
##  Median :0.0000   Median :  1400   Median :  8.00   Median :  3.000
##  Mean   :0.4934   Mean   :  3395   Mean   : 10.88   Mean   :  3.294
##  3rd Qu.:1.0000   3rd Qu.:  2800   3rd Qu.: 14.00   3rd Qu.:  4.000
##  Max.   :1.0000   Max.   :843300   Max.   :304.00   Max.   :116.000
##   num_keywords    data_channel_is_entertainment data_channel_is_bus
##  Min.   : 1.000   Min.   :0.000                 Min.   :0.0000
##  1st Qu.: 6.000   1st Qu.:0.000                 1st Qu.:0.0000
##  Median : 7.000   Median :0.000                 Median :0.0000
##  Mean   : 7.224   Mean   :0.178                 Mean   :0.1579
##  3rd Qu.: 9.000   3rd Qu.:0.000                 3rd Qu.:0.0000
##  Max.   :10.000   Max.   :1.000                 Max.   :1.0000
##  data_channel_is_socmed data_channel_is_tech weekday_is_tuesday
##  Min.   :0.0000         Min.   :0.0000       Min.   :0.0000
##  1st Qu.:0.0000         1st Qu.:0.0000       1st Qu.:0.0000
##  Median :0.0000         Median :0.0000       Median :0.0000
##  Mean   :0.0586         Mean   :0.1853       Mean   :0.1864
##  3rd Qu.:0.0000         3rd Qu.:0.0000       3rd Qu.:0.0000
##  Max.   :1.0000         Max.   :1.0000       Max.   :1.0000
##  weekday_is_wednesday weekday_is_friday weekday_is_saturday
##  Min.   :0.0000       Min.   :0.0000    Min.   :0.00000
##  1st Qu.:0.0000       1st Qu.:0.0000    1st Qu.:0.00000
##  Median :0.0000       Median :0.0000    Median :0.00000
##  Mean   :0.1875       Mean   :0.1438    Mean   :0.06188
```

```
##  3rd Qu.:0.0000       3rd Qu.:0.0000    3rd Qu.:0.00000
##  Max.   :1.0000       Max.   :1.0000    Max.   :1.00000
##    is_weekend     global_subjectivity min_positive_polarity
##  Min.   :0.0000   Min.   :0.0000      Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.3962      1st Qu.:0.05000
##  Median :0.0000   Median :0.4535      Median :0.10000
##  Mean   :0.1309   Mean   :0.4434      Mean   :0.09545
##  3rd Qu.:0.0000   3rd Qu.:0.5083      3rd Qu.:0.10000
##  Max.   :1.0000   Max.   :1.0000      Max.   :1.00000
##  title_subjectivity title_sentiment_polarity abs_title_subjectivity
##  Min.   :0.0000     Min.   :-1.00000         Min.   :0.0000
##  1st Qu.:0.0000     1st Qu.: 0.00000         1st Qu.:0.1667
##  Median :0.1500     Median : 0.00000         Median :0.5000
##  Mean   :0.2824     Mean   : 0.07143         Mean   :0.3418
##  3rd Qu.:0.5000     3rd Qu.: 0.15000         3rd Qu.:0.5000
##  Max.   :1.0000     Max.   : 1.00000         Max.   :0.5000
```

```
#put it into data frame bc why not...
sum_stats <- data.frame(summary(clean_data))
sum_stats
```

```
##     Var1                          Var2              Freq
## 1                     popular_article   Min.   :0.0000
## 2                     popular_article   1st Qu.:0.0000
## 3                     popular_article   Median :0.0000
## 4                     popular_article   Mean   :0.4934
## 5                     popular_article   3rd Qu.:1.0000
## 6                     popular_article   Max.   :1.0000
## 7                              shares   Min.   :     1
## 8                              shares   1st Qu.:   946
## 9                              shares   Median :  1400
## 10                             shares   Mean   :  3395
## 11                             shares   3rd Qu.:  2800
## 12                             shares   Max.   :843300
## 13                          num_hrefs   Min.   :  0.00
## 14                          num_hrefs   1st Qu.:  4.00
## 15                          num_hrefs   Median :  8.00
## 16                          num_hrefs   Mean   : 10.88
## 17                          num_hrefs   3rd Qu.: 14.00
## 18                          num_hrefs   Max.   :304.00
## 19                     num_self_hrefs   Min.   :  0.000
## 20                     num_self_hrefs   1st Qu.:  1.000
## 21                     num_self_hrefs   Median :  3.000
## 22                     num_self_hrefs   Mean   :  3.294
## 23                     num_self_hrefs   3rd Qu.:  4.000
## 24                     num_self_hrefs   Max.   :116.000
## 25                       num_keywords   Min.   : 1.000
## 26                       num_keywords   1st Qu.: 6.000
## 27                       num_keywords   Median : 7.000
## 28                       num_keywords   Mean   : 7.224
## 29                       num_keywords   3rd Qu.: 9.000
## 30                       num_keywords   Max.   :10.000
## 31       data_channel_is_entertainment    Min.   :0.000
## 32       data_channel_is_entertainment    1st Qu.:0.000
## 33       data_channel_is_entertainment    Median :0.000
```

```
## 34       data_channel_is_entertainment    Mean   :0.178
## 35       data_channel_is_entertainment    3rd Qu.:0.000
## 36       data_channel_is_entertainment    Max.   :1.000
## 37                 data_channel_is_bus   Min.   :0.0000
## 38                 data_channel_is_bus   1st Qu.:0.0000
## 39                 data_channel_is_bus   Median :0.0000
## 40                 data_channel_is_bus   Mean   :0.1579
## 41                 data_channel_is_bus   3rd Qu.:0.0000
## 42                 data_channel_is_bus   Max.   :1.0000
## 43              data_channel_is_socmed   Min.   :0.0000
## 44              data_channel_is_socmed   1st Qu.:0.0000
## 45              data_channel_is_socmed   Median :0.0000
## 46              data_channel_is_socmed   Mean   :0.0586
## 47              data_channel_is_socmed   3rd Qu.:0.0000
## 48              data_channel_is_socmed   Max.   :1.0000
## 49                data_channel_is_tech   Min.   :0.0000
## 50                data_channel_is_tech   1st Qu.:0.0000
## 51                data_channel_is_tech   Median :0.0000
## 52                data_channel_is_tech   Mean   :0.1853
## 53                data_channel_is_tech   3rd Qu.:0.0000
## 54                data_channel_is_tech   Max.   :1.0000
## 55                  weekday_is_tuesday   Min.   :0.0000
## 56                  weekday_is_tuesday   1st Qu.:0.0000
## 57                  weekday_is_tuesday   Median :0.0000
## 58                  weekday_is_tuesday   Mean   :0.1864
## 59                  weekday_is_tuesday   3rd Qu.:0.0000
## 60                  weekday_is_tuesday   Max.   :1.0000
## 61                weekday_is_wednesday   Min.   :0.0000
## 62                weekday_is_wednesday   1st Qu.:0.0000
## 63                weekday_is_wednesday   Median :0.0000
## 64                weekday_is_wednesday   Mean   :0.1875
## 65                weekday_is_wednesday   3rd Qu.:0.0000
## 66                weekday_is_wednesday   Max.   :1.0000
## 67                   weekday_is_friday   Min.   :0.0000
## 68                   weekday_is_friday   1st Qu.:0.0000
## 69                   weekday_is_friday   Median :0.0000
## 70                   weekday_is_friday   Mean   :0.1438
## 71                   weekday_is_friday   3rd Qu.:0.0000
## 72                   weekday_is_friday   Max.   :1.0000
## 73                 weekday_is_saturday  Min.   :0.00000
## 74                 weekday_is_saturday  1st Qu.:0.00000
## 75                 weekday_is_saturday  Median :0.00000
## 76                 weekday_is_saturday  Mean   :0.06188
## 77                 weekday_is_saturday  3rd Qu.:0.00000
## 78                 weekday_is_saturday  Max.   :1.00000
## 79                           is_weekend   Min.   :0.0000
## 80                           is_weekend   1st Qu.:0.0000
## 81                           is_weekend   Median :0.0000
## 82                           is_weekend   Mean   :0.1309
## 83                           is_weekend   3rd Qu.:0.0000
## 84                           is_weekend   Max.   :1.0000
## 85                 global_subjectivity   Min.   :0.0000
## 86                 global_subjectivity   1st Qu.:0.3962
## 87                 global_subjectivity   Median :0.4535
```

```
## 88                 global_subjectivity   Mean   :0.4434
## 89                 global_subjectivity   3rd Qu.:0.5083
## 90                 global_subjectivity   Max.   :1.0000
## 91              min_positive_polarity  Min.   :0.00000
## 92              min_positive_polarity  1st Qu.:0.05000
## 93              min_positive_polarity  Median :0.10000
## 94              min_positive_polarity  Mean   :0.09545
## 95              min_positive_polarity  3rd Qu.:0.10000
## 96              min_positive_polarity  Max.   :1.00000
## 97                 title_subjectivity   Min.   :0.0000
## 98                 title_subjectivity   1st Qu.:0.0000
## 99                 title_subjectivity   Median :0.1500
## 100                title_subjectivity   Mean   :0.2824
## 101                title_subjectivity   3rd Qu.:0.5000
## 102                title_subjectivity   Max.   :1.0000
## 103          title_sentiment_polarity Min.   :-1.00000
## 104          title_sentiment_polarity 1st Qu.: 0.00000
## 105          title_sentiment_polarity Median : 0.00000
## 106          title_sentiment_polarity Mean   : 0.07143
## 107          title_sentiment_polarity 3rd Qu.: 0.15000
## 108          title_sentiment_polarity Max.   : 1.00000
## 109            abs_title_subjectivity   Min.   :0.0000
## 110            abs_title_subjectivity   1st Qu.:0.1667
## 111            abs_title_subjectivity   Median :0.5000
## 112            abs_title_subjectivity   Mean   :0.3418
## 113            abs_title_subjectivity   3rd Qu.:0.5000
## 114            abs_title_subjectivity   Max.   :0.5000
```

```
#Group-by function to spit out summary stats of our selected variables...grouping by popular (1) vs unp
by_popular <- clean_data %>% group_by(popular_article)
by_popular <- by_popular %>% summarise_all(list(min = min,
                                                mean = mean,
                                                median = median,
                                                max = max,
                                                sd = sd), na.rm = TRUE)
by_popular
```

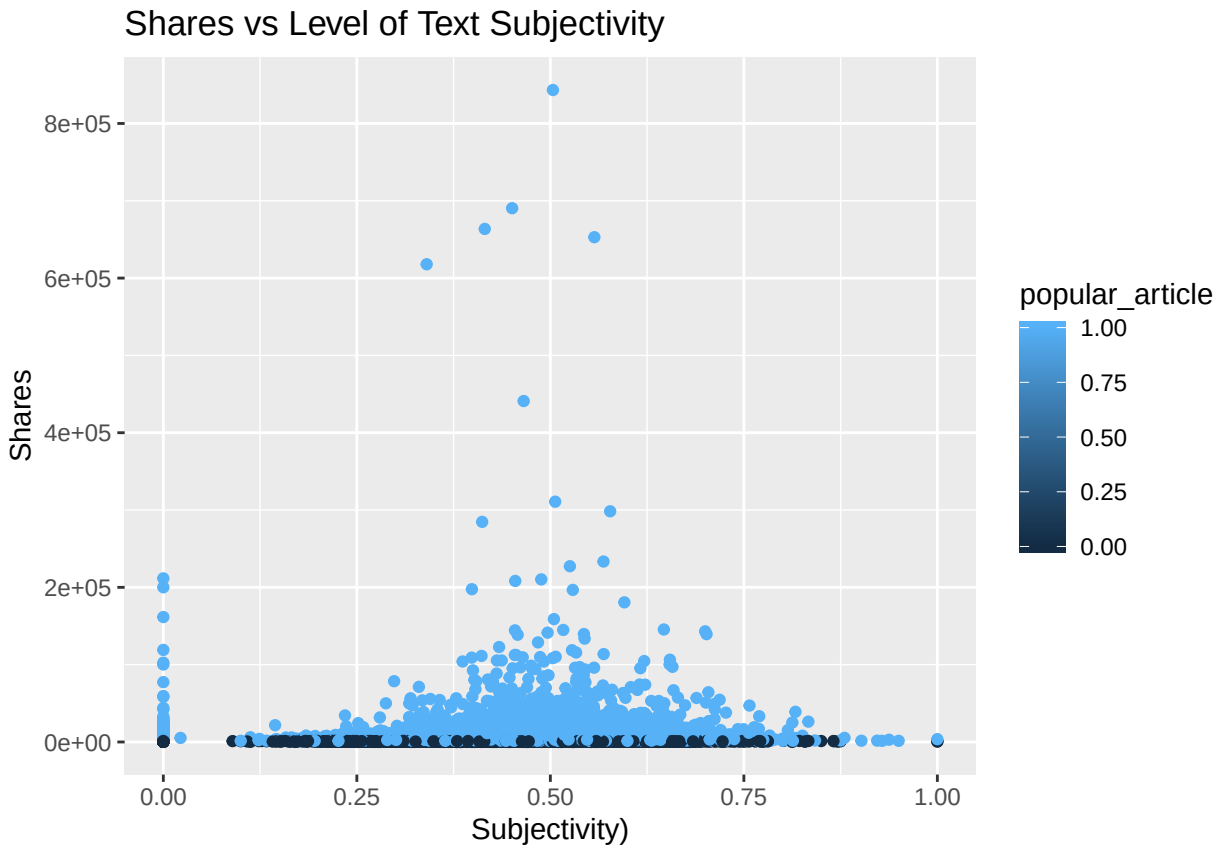
```
## # A tibble: 2 x 91
##   popular_article shares_min num_hrefs_min num_self_hrefs_~
##             <dbl>      <int>         <int>            <int>
## 1               0          1             0                0
## 2               1       1500             0                0
## # ... with 87 more variables: num_keywords_min <int>,
## #   data_channel_is_entertainment_min <int>,
## #   data_channel_is_bus_min <int>, data_channel_is_socmed_min <int>,
## #   data_channel_is_tech_min <int>, weekday_is_tuesday_min <int>,
## #   weekday_is_wednesday_min <int>, weekday_is_friday_min <int>,
## #   weekday_is_saturday_min <int>, is_weekend_min <int>,
## #   global_subjectivity_min <dbl>, min_positive_polarity_min <dbl>,
## #   title_subjectivity_min <dbl>, title_sentiment_polarity_min <dbl>,
## #   abs_title_subjectivity_min <dbl>, shares_mean <dbl>,
## #   num_hrefs_mean <dbl>, num_self_hrefs_mean <dbl>,
## #   num_keywords_mean <dbl>, data_channel_is_entertainment_mean <dbl>,
## #   data_channel_is_bus_mean <dbl>, data_channel_is_socmed_mean <dbl>,
## #   data_channel_is_tech_mean <dbl>, weekday_is_tuesday_mean <dbl>,
```

```
## #   weekday_is_wednesday_mean <dbl>, weekday_is_friday_mean <dbl>,
## #   weekday_is_saturday_mean <dbl>, is_weekend_mean <dbl>,
## #   global_subjectivity_mean <dbl>, min_positive_polarity_mean <dbl>,
## #   title_subjectivity_mean <dbl>, title_sentiment_polarity_mean <dbl>,
## #   abs_title_subjectivity_mean <dbl>, shares_median <dbl>,
## #   num_hrefs_median <dbl>, num_self_hrefs_median <dbl>,
## #   num_keywords_median <dbl>, data_channel_is_entertainment_median <dbl>,
## #   data_channel_is_bus_median <dbl>, data_channel_is_socmed_median <dbl>,
## #   data_channel_is_tech_median <dbl>, weekday_is_tuesday_median <dbl>,
## #   weekday_is_wednesday_median <dbl>, weekday_is_friday_median <dbl>,
## #   weekday_is_saturday_median <dbl>, is_weekend_median <dbl>,
## #   global_subjectivity_median <dbl>, min_positive_polarity_median <dbl>,
## #   title_subjectivity_median <dbl>,
## #   title_sentiment_polarity_median <dbl>,
## #   abs_title_subjectivity_median <dbl>, shares_max <int>,
## #   num_hrefs_max <int>, num_self_hrefs_max <int>, num_keywords_max <int>,
## #   data_channel_is_entertainment_max <int>,
## #   data_channel_is_bus_max <int>, data_channel_is_socmed_max <int>,
## #   data_channel_is_tech_max <int>, weekday_is_tuesday_max <int>,
## #   weekday_is_wednesday_max <int>, weekday_is_friday_max <int>,
## #   weekday_is_saturday_max <int>, is_weekend_max <int>,
## #   global_subjectivity_max <dbl>, min_positive_polarity_max <dbl>,
## #   title_subjectivity_max <dbl>, title_sentiment_polarity_max <dbl>,
## #   abs_title_subjectivity_max <dbl>, shares_sd <dbl>, num_hrefs_sd <dbl>,
## #   num_self_hrefs_sd <dbl>, num_keywords_sd <dbl>,
## #   data_channel_is_entertainment_sd <dbl>, data_channel_is_bus_sd <dbl>,
## #   data_channel_is_socmed_sd <dbl>, data_channel_is_tech_sd <dbl>,
## #   weekday_is_tuesday_sd <dbl>, weekday_is_wednesday_sd <dbl>,
## #   weekday_is_friday_sd <dbl>, weekday_is_saturday_sd <dbl>,
## #   is_weekend_sd <dbl>, global_subjectivity_sd <dbl>,
## #   min_positive_polarity_sd <dbl>, title_subjectivity_sd <dbl>,
## #   title_sentiment_polarity_sd <dbl>, abs_title_subjectivity_sd <dbl>
```

```r
#test and training split pt.2
set.seed(1861)
train_idx <- sample(1:nrow(clean_data), size = floor(nrow(clean_data) * .75))
mash_train <- clean_data %>% slice(train_idx)
mash_test <- clean_data %>% slice(-train_idx)
```

Scatter of shares vs global subjectivity

```r
p1 <- ggplot(mash_train, aes(x = global_subjectivity, y = shares)) + geom_point(mapping = aes(color = p
  labs(x = "Subjectivity)", y = "Shares", title = "Shares vs Level of Text Subjectivity")
plot(p1)
```




```
#This is interesting bc the articles with the most amount of shares appear to have the highest concetra
```

Ridgline Density

```
library(ggridges)
```

```
##
## Attaching package: 'ggridges'
```

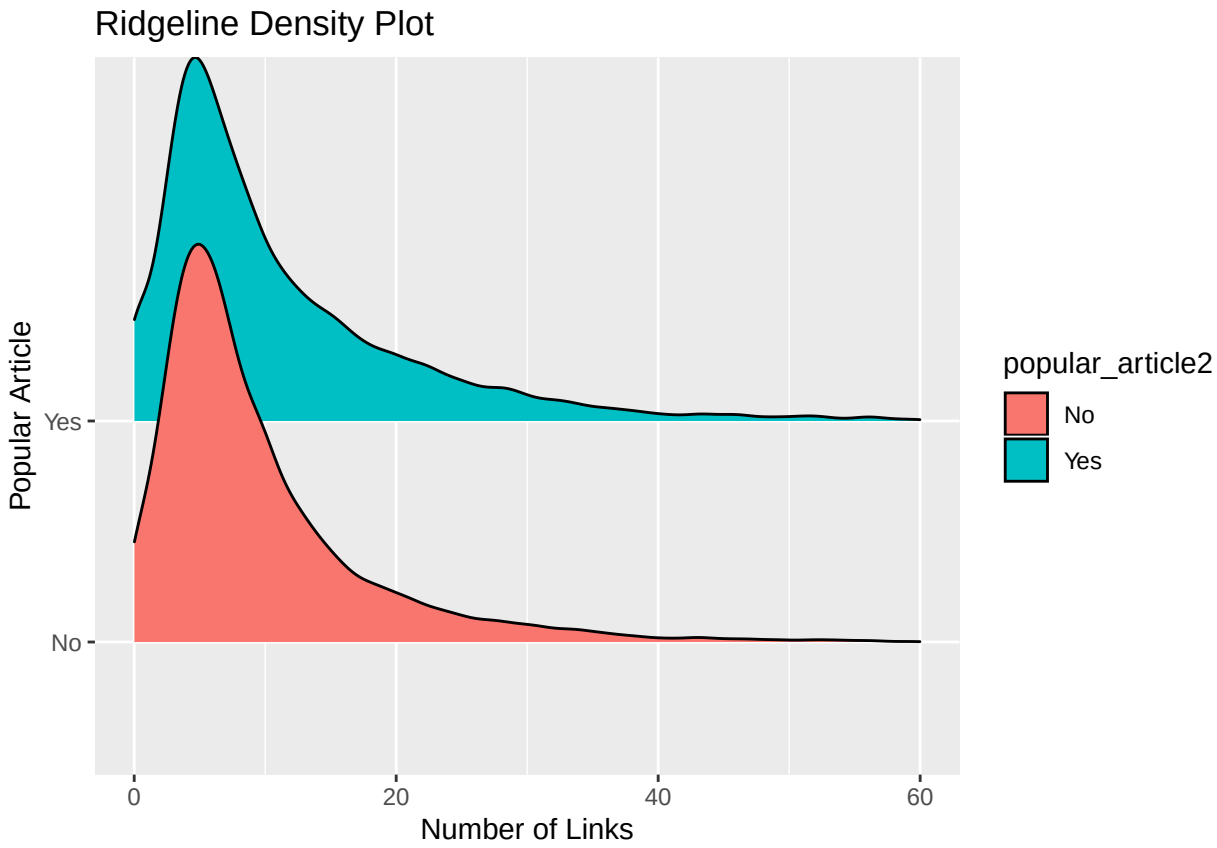
```
## The following object is masked from 'package:ggplot2':
##
##     scale_discrete_manual
```

```
data_forplot <- mash_train %>% mutate(popular_article2 = ifelse( popular_article == 1,"Yes",
                                                                 "No"))

p2 <- ggplot(data_forplot, aes(x = num_hrefs, y = popular_article2, fill = popular_article2)) +
  geom_density_ridges() +
  scale_x_continuous(limits = c(0,60)) +
  labs(x = "Number of Links", y = "Popular Article", title = "Ridgeline Density Plot")
plot(p2)
```

```
## Picking joint bandwidth of 0.936
```

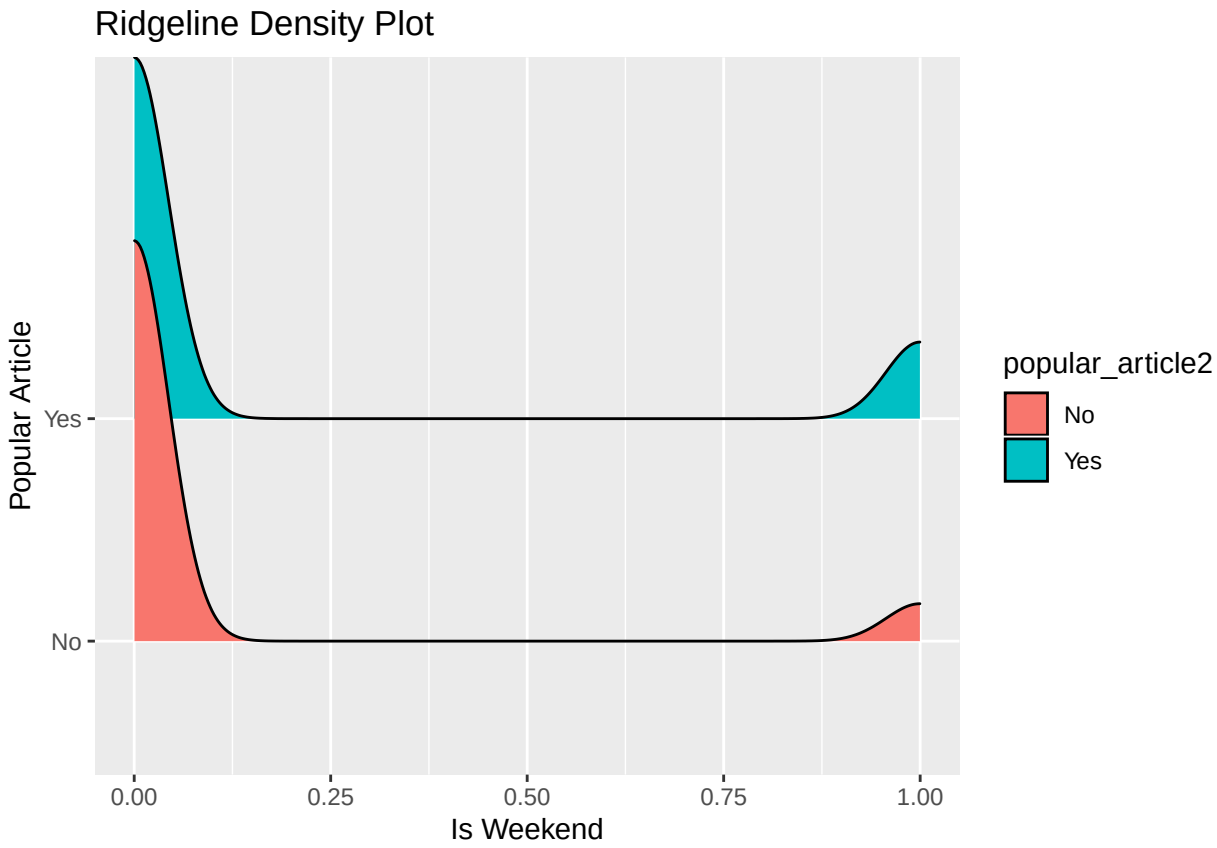
```
## Warning: Removed 230 rows containing non-finite values
## (stat_density_ridges).
```




```
#This plot seems to show that the more popular articles have a higher number of links on them, for ad p
#Needs further investigation!

p3 <- ggplot(data_forplot, aes(x = is_weekend, y = popular_article2, fill = popular_article2)) +
  geom_density_ridges() +
  scale_x_continuous(limits = c(0,1)) +
  labs(x = "Is Weekend", y = "Popular Article", title = "Ridgeline Density Plot")
plot(p3)
```

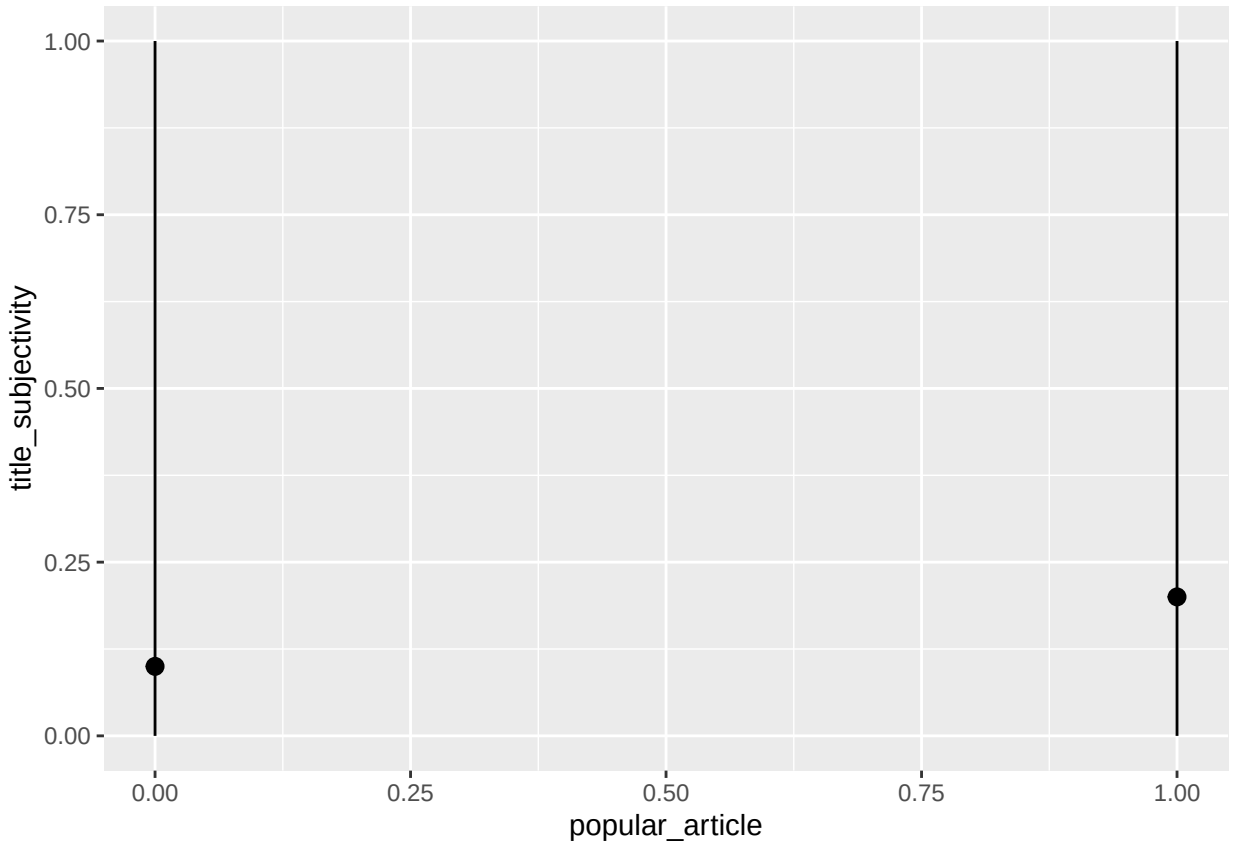
```
## Picking joint bandwidth of 0.0435
```




```
#This plot is interesting because it appears that a larger proportion of popular articles had a greater
```

Stat Summary

```
p4 <- ggplot(data = mash_train) +
  stat_summary(
    mapping = aes(x = popular_article, y = title_subjectivity),
    fun.ymin = min,
    fun.ymax = max,
    fun.y = median
  )
plot(p4)
```

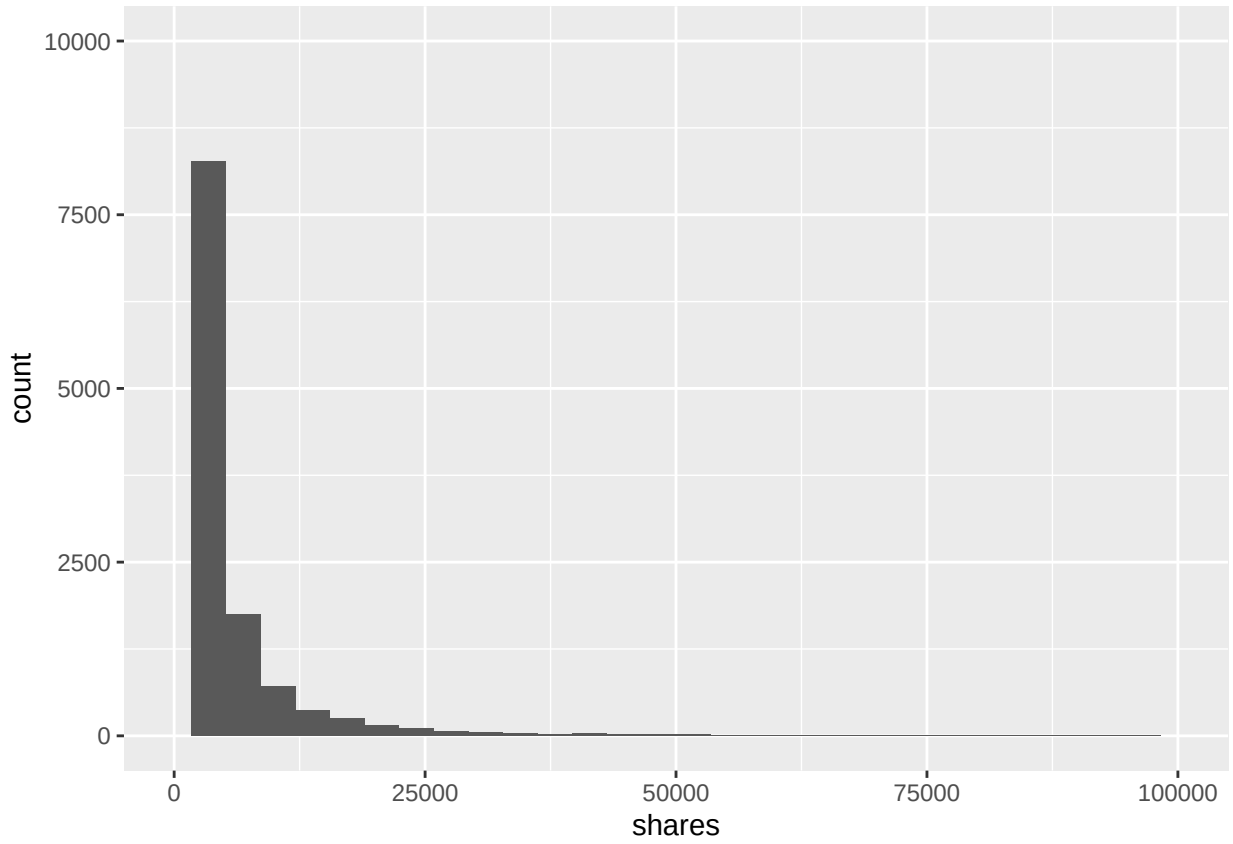



```
#This plot highlights the difference in title subjectivity (personal sentiment level of title) between
```

Histogram/ Distribution of Shares Data

```
p5 <- ggplot(data = mash_train, aes(x=shares)) + geom_histogram() + scale_x_continuous(limits = c(0,100
  scale_y_continuous(limits = c(0,10000))
plot(p5)
```

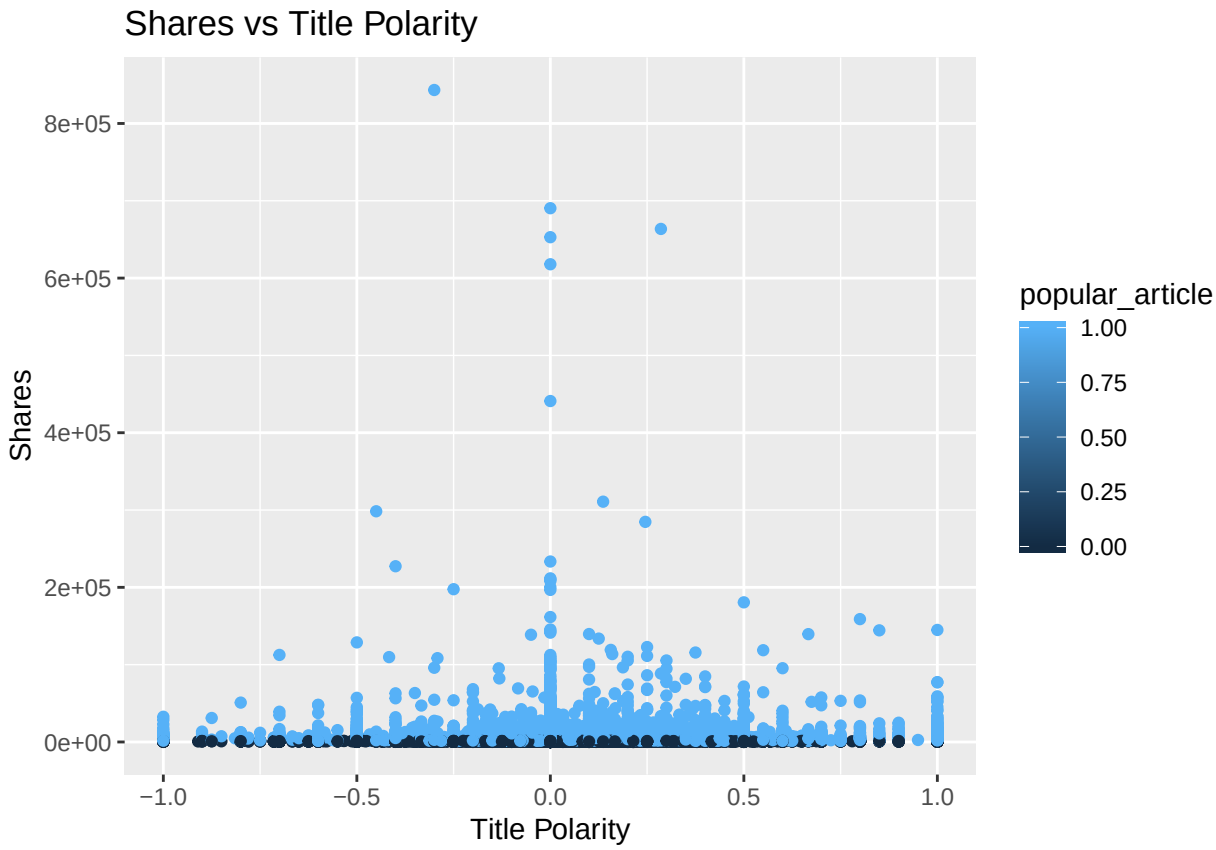
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 52 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
#Here we hope to highlight the fact that the distribution of shares is not normally distributed. That i
```

Scatter Shares vs Title Polarity

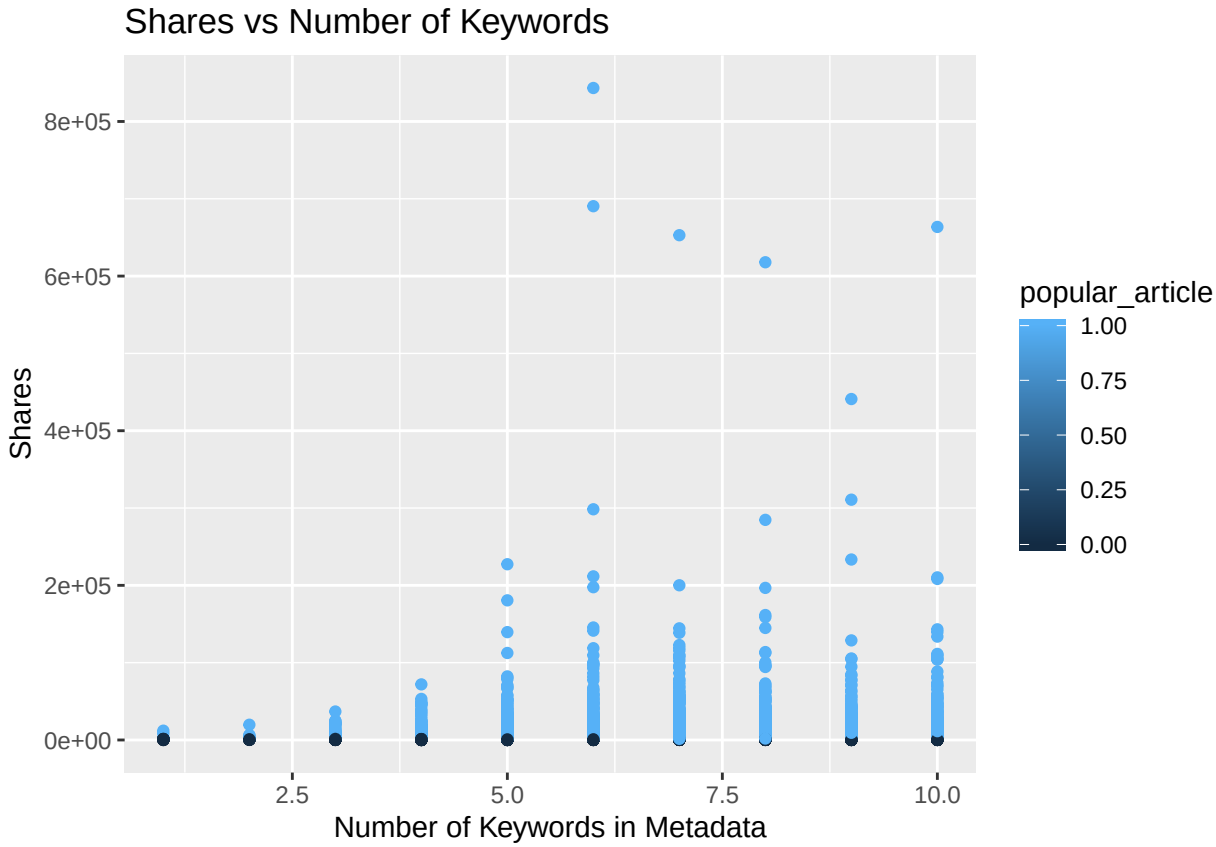
```
p6 <- ggplot(mash_train, aes(x = title_sentiment_polarity, y = shares)) + geom_point(mapping = aes(colo
  labs(x = "Title Polarity", y = "Shares", title = "Shares vs Title Polarity")
plot(p6)
```




```
#This plot is interesting because it shows that articles with neutral titles (title polarity of 0) seem
```

Scatter of Shares vs Number of Keywords

```
p7 <- ggplot(mash_train, aes(x = num_keywords, y = shares)) + geom_point(mapping = aes(color = popular_
  labs(x = "Number of Keywords in Metadata", y = "Shares", title = "Shares vs Number of Keywords")
plot(p7)
```




```
#This plot is interesting because it seems that articles with more than 5 keywords have more shares and
```

The plots above and their variables were selected based off of the summary statistics we ran because we wanted to highlight what we might find to be the key differences/key variables in predicting either the number of shares (OLS model) or whether or not an article would be popular or not (logistic model).

Further, we wanted to create a historgram/simple count of our number of shares to visualize the distrinution of the data. It appears to be skewed to the right. So, if we do decide to clean the data further and remove outliers, we may not be able to utilize z-scores because the data is not normally distributed. If we do remove outliers then we will have to run our variable selection, VIF, and OLS models again.