Viewing Trends in Media Articles

Patrick C
pcosta

Due Wed, March 24, at 8:00PM (Pittburgh time)

Contents

```
library("knitr")
library("cmu202")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("jtools")
library("leaps")
```

```
social <- readr::read_csv("http://stat.cmu.edu/~gordonw/social.csv")
```

Introduction

As technology continues to evolve at a rapid pace, numerous new aspects of human culture and habits are changing as well. Within the past 10-15 years, the invention and populzarization of social media and media websites has resulted in a large shift in how people (particularly younger generations) interact with the internet. With this increased popularity and userbase, it is important to view trends on said platforms through meaningful data analysis. In this paper, we will focus on the number of shares an article has when considering other factor(s) that result in such an action.

Exploratory Data Analysis

In this data, we analyze a sample of 388 unique articles with 4 variables from collection from published Mashable articles over a two year period. As our ultimate goal is to view popularity trends on media platforms, we examine the relationship between number of article shares compared to our three explanatory variables (content, images, and day published). The variables can be summarized as follows:

Shares: The number of shares the article has received (# shares). Content: The number of words in the article (# words). Images: The number of images in the article (# images). Day Published: The day of the week on which the article was published (Monday,. . . ,Sunday) (day).

The first few lines of the data table are as follows.

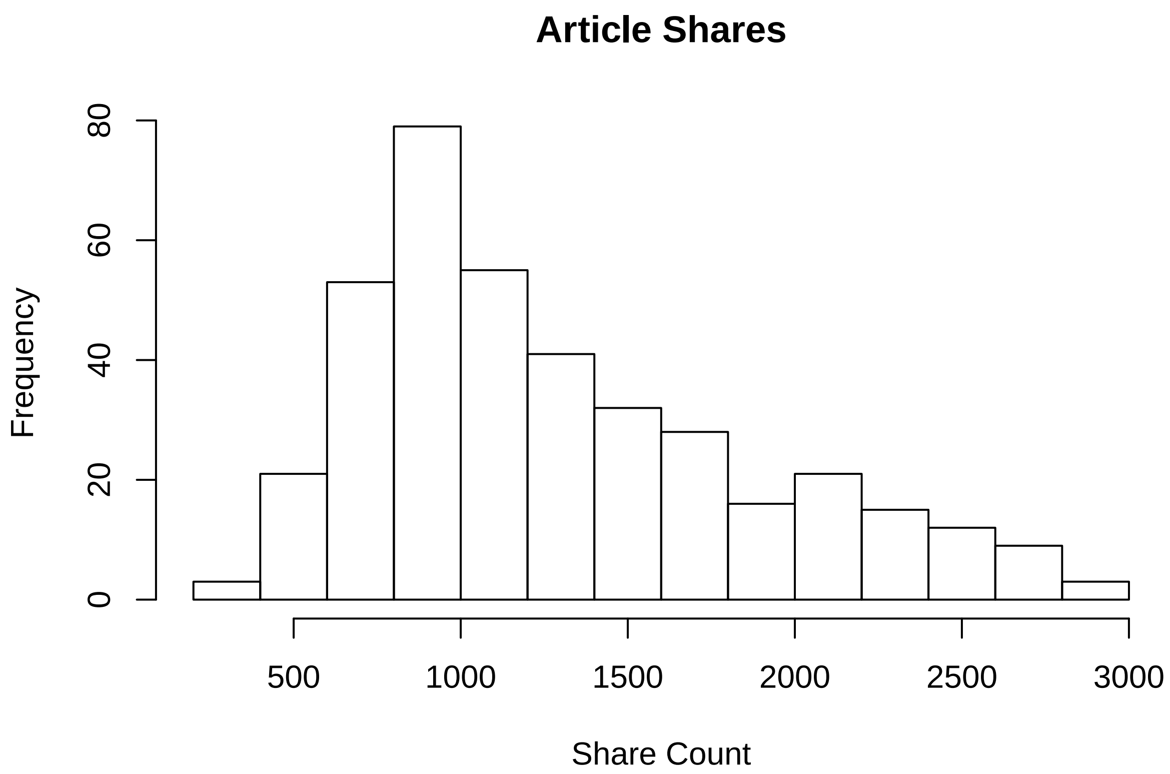
```
head(social)
```

```
## # A tibble: 6 x 4
##   shares content images daypublished
##    <dbl>   <dbl>  <dbl> <chr>
## 1   1100     367      1 Monday
## 2   1400     712      1 Monday
## 3    479     291      1 Monday
## 4   2500     463      5 Monday
## 5   1200     498     13 Monday
## 6   1200    1084      1 Monday
```

Univariate Exploration

To begin univariate analysis of our data, we create and observe a series of histrograms and bar plots based on the all four measured variables within the dataset. Each chart also includes a numerical summary as well. **For Shares:**

```
hist(social$shares,
     main = 'Article Shares',
     xlab = 'Share Count')
```

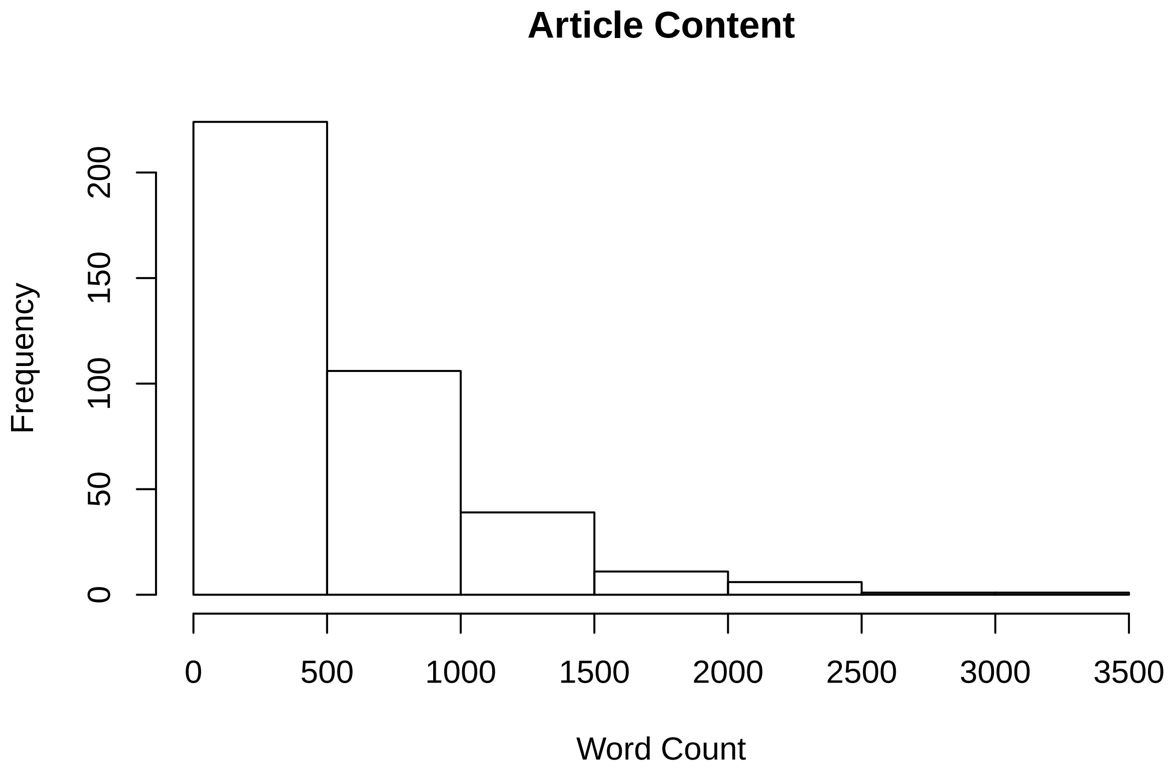


```r
summary(social$shares)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   319.0   859.8  1200.0  1325.1  1700.0  2900.0
```

For Content:

```r
hist(social$content,
     main = 'Article Content',
     xlab = 'Word Count')
```

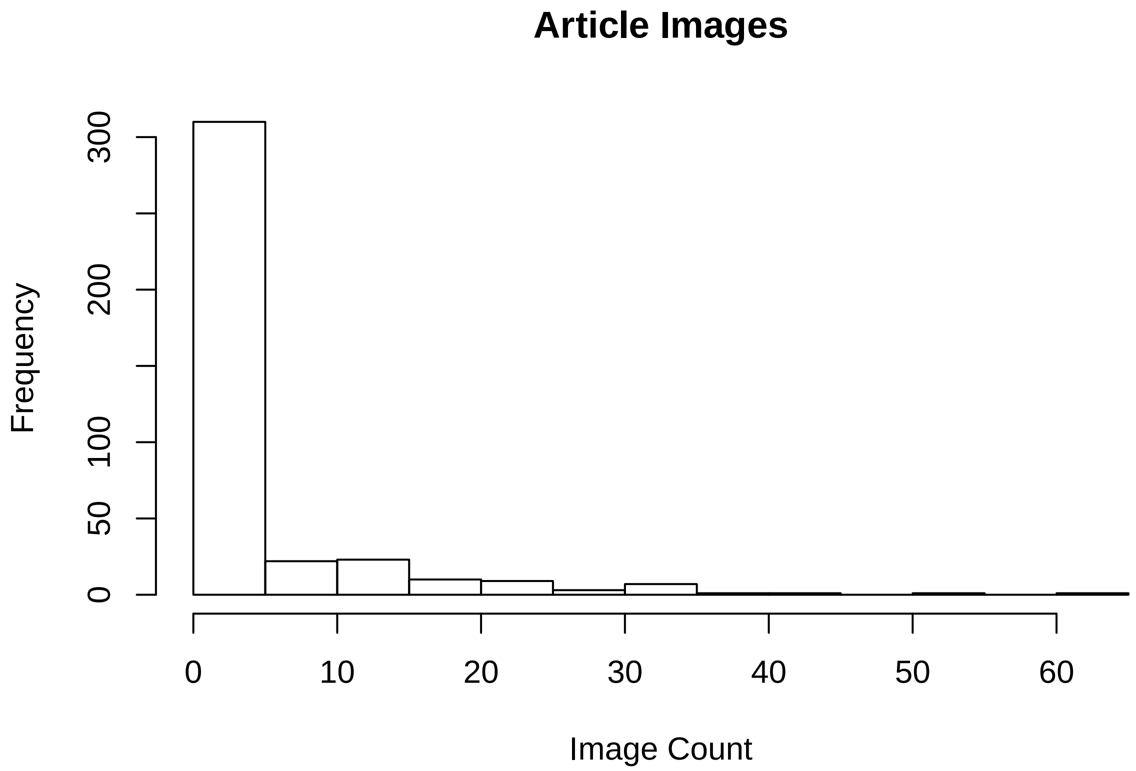


```r
summary(social$content)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   276.5   433.0   586.1   734.2  3174.0
```

For Images:

```r
hist(social$images,
     main = 'Article Images',
     xlab = 'Image Count')
```

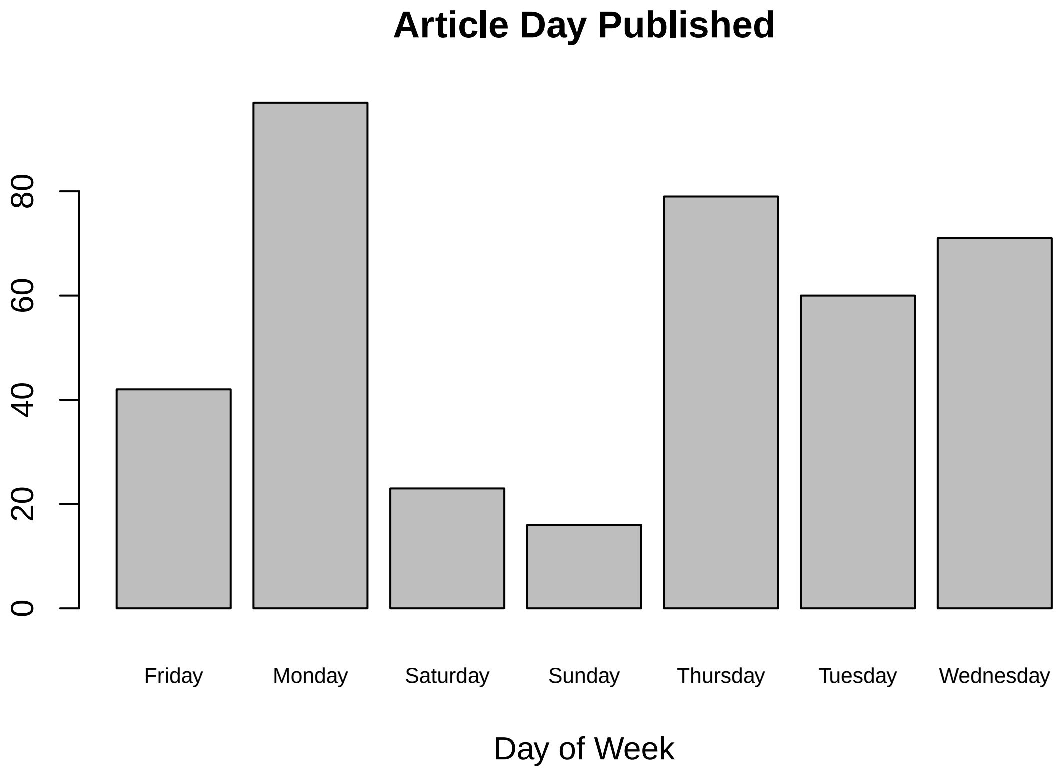


```
summary(social$images)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.000   1.000   4.433   3.000  61.000
```

For Day Published:

```
tempTable <- table(social$daypublished)
barplot(tempTable,
     main = 'Article Day Published',
     xlab = 'Day of Week',
     cex.names = 0.7)
```



```
summary(social$daypublished)
```

```
##    Length     Class      Mode 
##       388 character character 
```

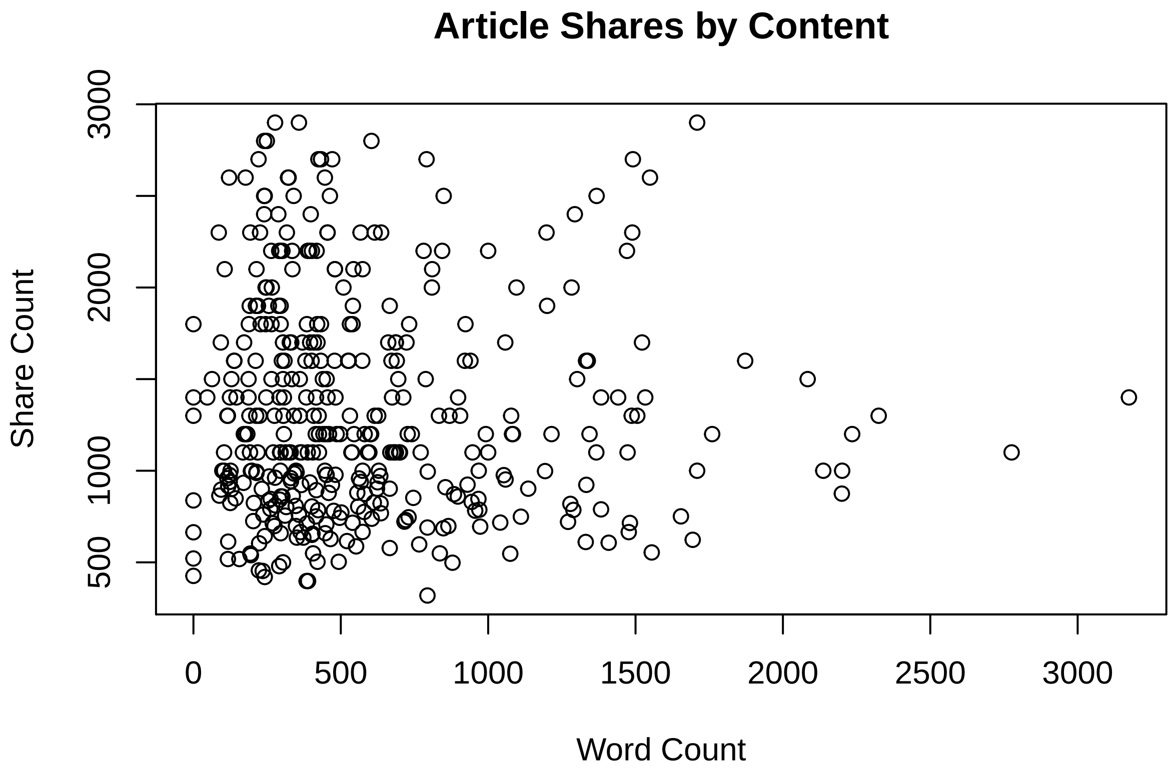
After viewing the chart visualizations and numerical summaries of all variables, we can make the following observations: The **article shares** distribution is unimodal and right-skewed, with the mean and median being relatively near each other at 1325 and 1200, respectively. There appears to be a wide range of share counts (values range from 319 to 2900 shares), with the largest spike being around 800-1000 shares for an article. The **article content** distribution is unimodal and strongly right-skewed. There still appears to be a wide range of word count, as values range from 0 to 3174 words. There is a sizeable spike in word count between 0 to 500 words and median word count of 433 words. The **article images** distribution is unimodal and extremely right skewed. Values range from 0 to 61, and there are potential outliers around 50 and 60 images. The chart also features a large spike at values between 0 and 5 images, coupled by a median of 1 image. Lastly, the **article day published** plot shows the most frequent publish days to be Monday and Thursday, with the least frequest days to be over the weekend.

Bivariate Exploration

Having looked at each individual variable within the dataset, we can now move on to bivariate EDA with each explanatory variable compared to shares.

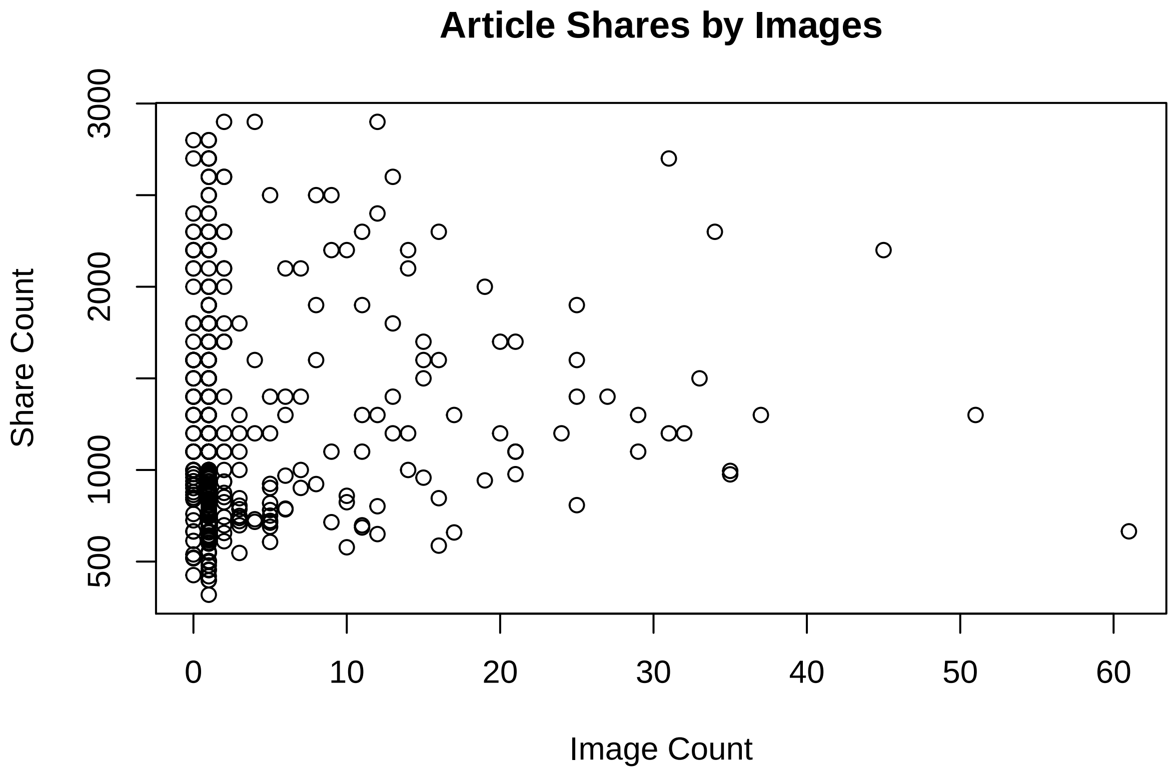
For Shares~Content:

```
plot(shares~content,
     data = social,
     main = 'Article Shares by Content',
     xlab = 'Word Count',
     ylab = 'Share Count')
```



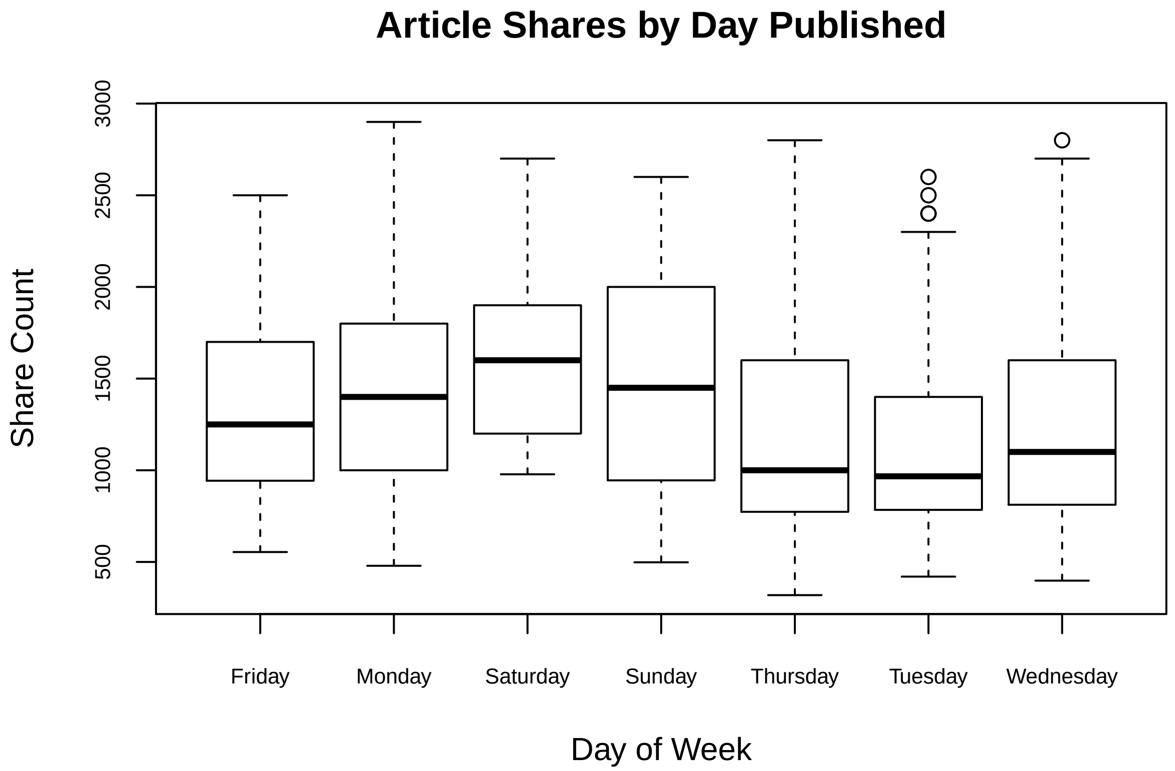
For Shares~Images:

```
plot(shares~images,
      data = social,
      main = 'Article Shares by Images',
      xlab = 'Image Count',
      ylab = 'Share Count')
```



For Shares~Day Published:

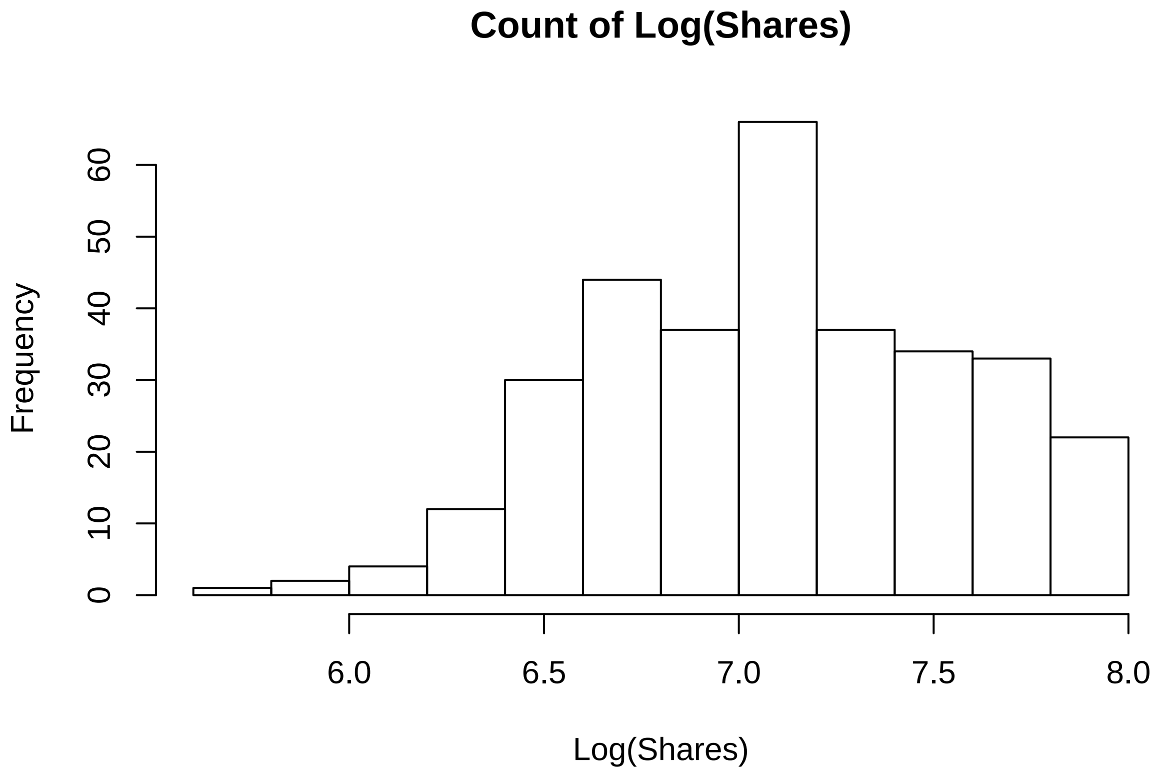
```
boxplot(shares~daypublished,
     data = social,
     main = 'Article Shares by Day Published',
     xlab = 'Day of Week',
     ylab = 'Share Count',
     cex.axis= 0.7)
```

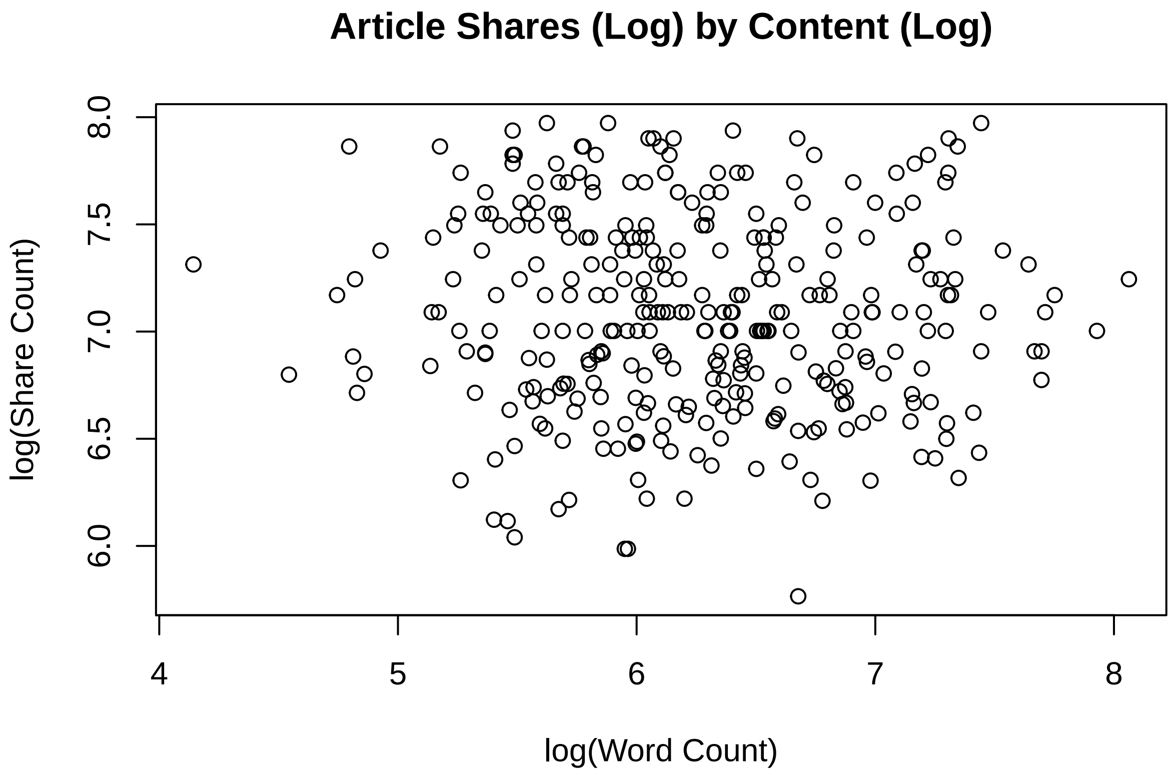


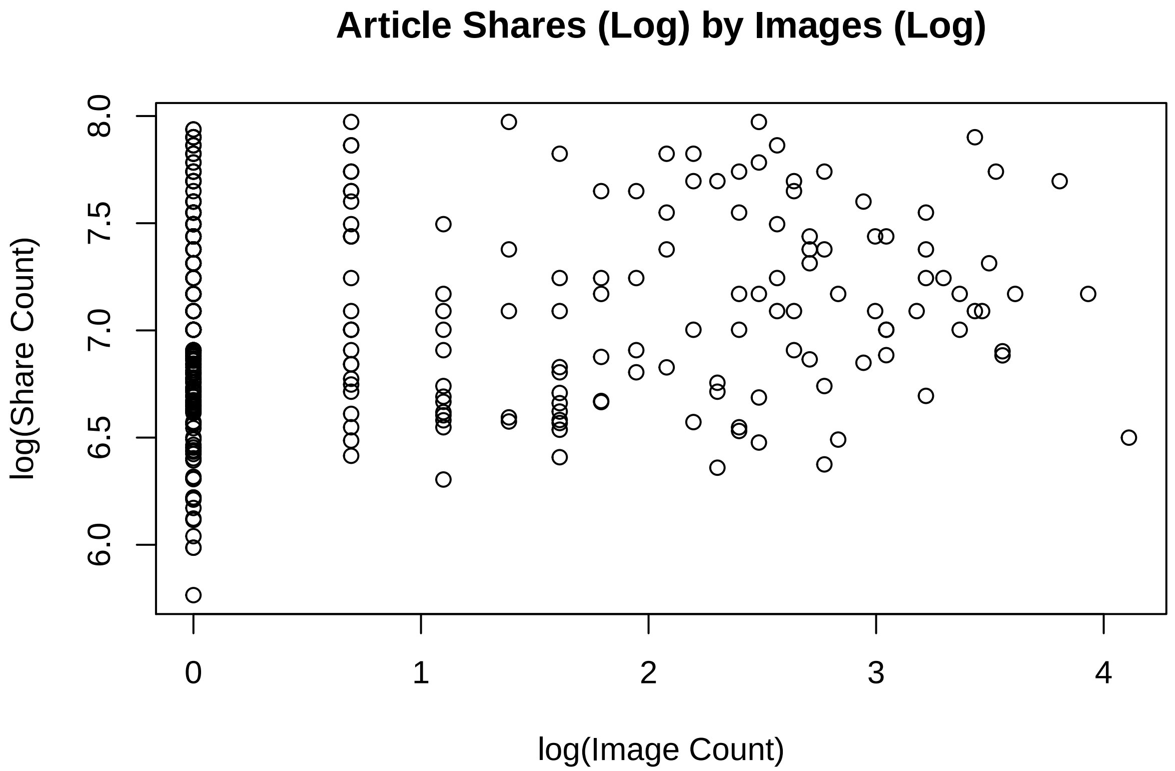
Analyzing our graphs, it is apparent that neither Image Count nor Word Count (content) have strong linear relationships with shares on their own (without transformations). Intead, both quantitative explanatory variables have a dense amount of points found closer to each respective chart's origin. Share count by Day Published reveals no noticeable differences between days, as all boxes overlap with eachother. Saturday features the highest median share count, while Tuesday and Wednesday have a few noted outliers.

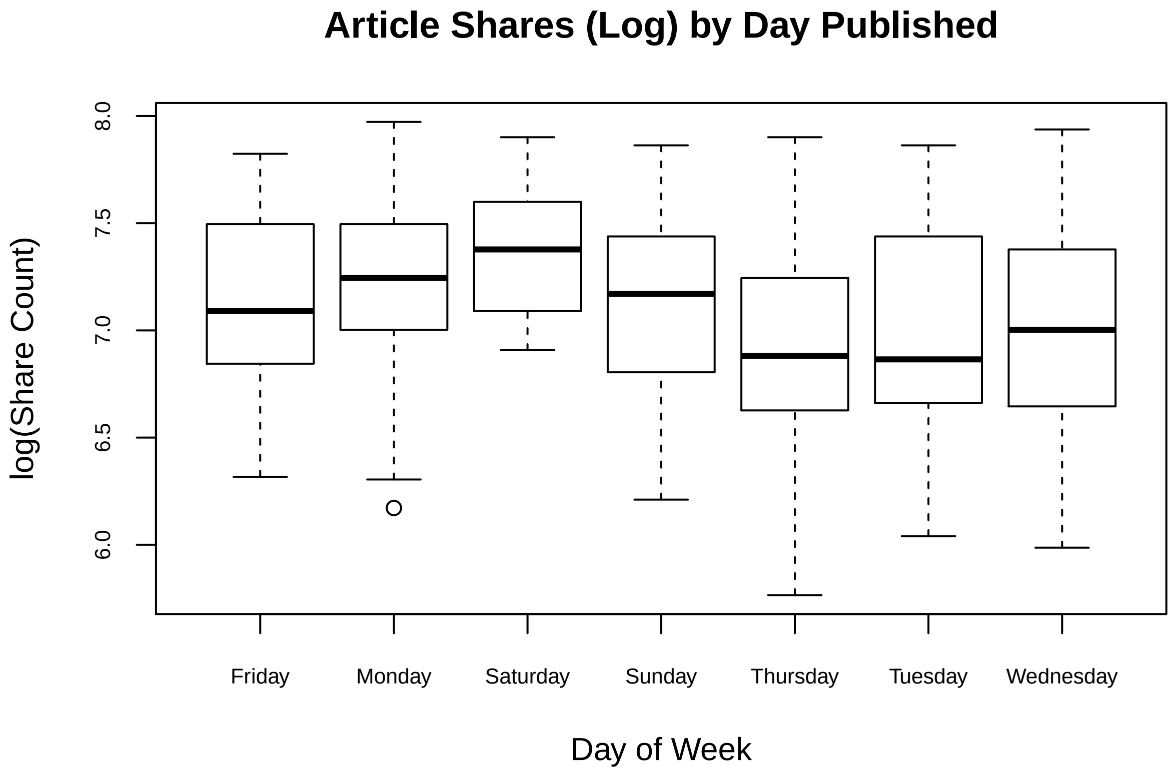
Modeling

We now look to create a linear regression model with given variables to understand and predict the number of shares for a given article. As the previous histogram of our response variable is right-skewed, we will perform a transformation on the **article shares** variable to create an even distribution of said chart. Our explanatory quantitative variables are strongly right-skewed as well, so more transformations will be performed to create more reasonably justified residual and qq plots of those variables when compared to article shares. To accomplish this, we use a series of logarithmic transformations and data cleaning to remove hazardous values to result in the following charts:


Count of Log(Shares)
Frequency
0
10
20
30
40
50
60
6.0
6.5
7.0
7.5
8.0
Log(Shares)


Article Shares (Log) by Content (Log)
log(Share Count)
6.0
6.5
7.0
7.5
8.0
4
5
6
7
8
log(Word Count)


Article Shares (Log) by Images (Log)
log(Share Count)
8.0
7.5
7.0
6.5
6.0
0
1
2
3
4
log(Image Count)



While the updated charts feature a more normal article share distribution and reasonable QQ/residual plots for log(shares)~log(content) and log(shares)~log(images), the linearity between log(shares)~log(content) and log(shares)~log(images) is not very strong. With the linearity assumption and residual plot hard to justify for log(shares)~log(images), we choose to remove this variable from the model in an effort to maintain simplicity

To check for dangerous multicollinearity between explanatory variables, we will observe GVIF values from the remaining variables.

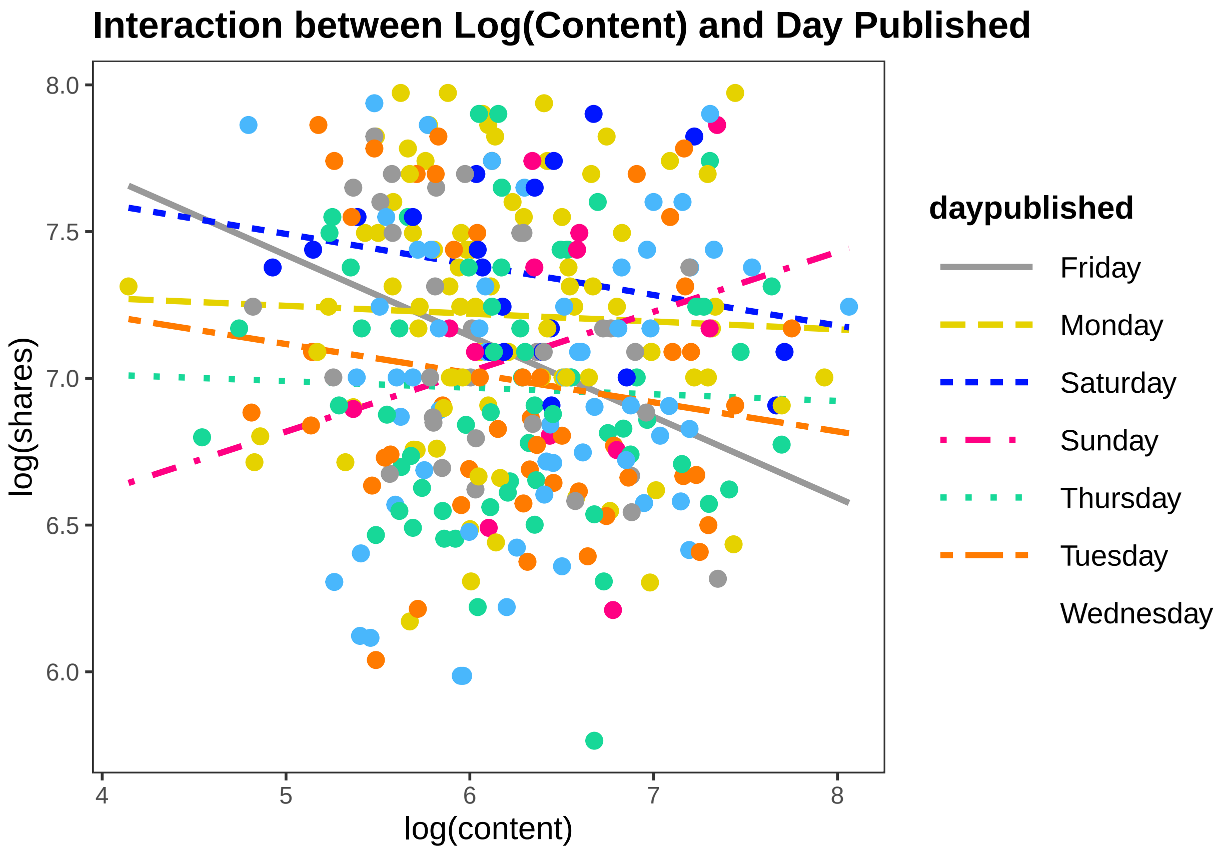
```
gvifTest.mod <- lm(log.shares ~ log.content + daypublished,
                    data = social.transform)
car::vif(gvifTest.mod)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## log.content  1.012367  1        1.006164
## daypublished 1.012367  6        1.001025
```

All variables feature GVIF values under our maximum allowed of 2.5, so we do not have issues of multicollinearity and will continue using the remaining variables within our model. We then turn to discuss potential interaction between log(content) and daypublished through the following chart and numerical summary:

```
interact.mod <- lm(log.shares ~ log.content + daypublished + log.content:daypublished,
               data = social.transform)

interact_plot(interact.mod,
              pred = log.content,
              modx = daypublished,
              y.label = "log(shares)",
              x.label = "log(content)",
              main = "Interaction between Log(Content) and Day Published",
              plot.points = TRUE)
```



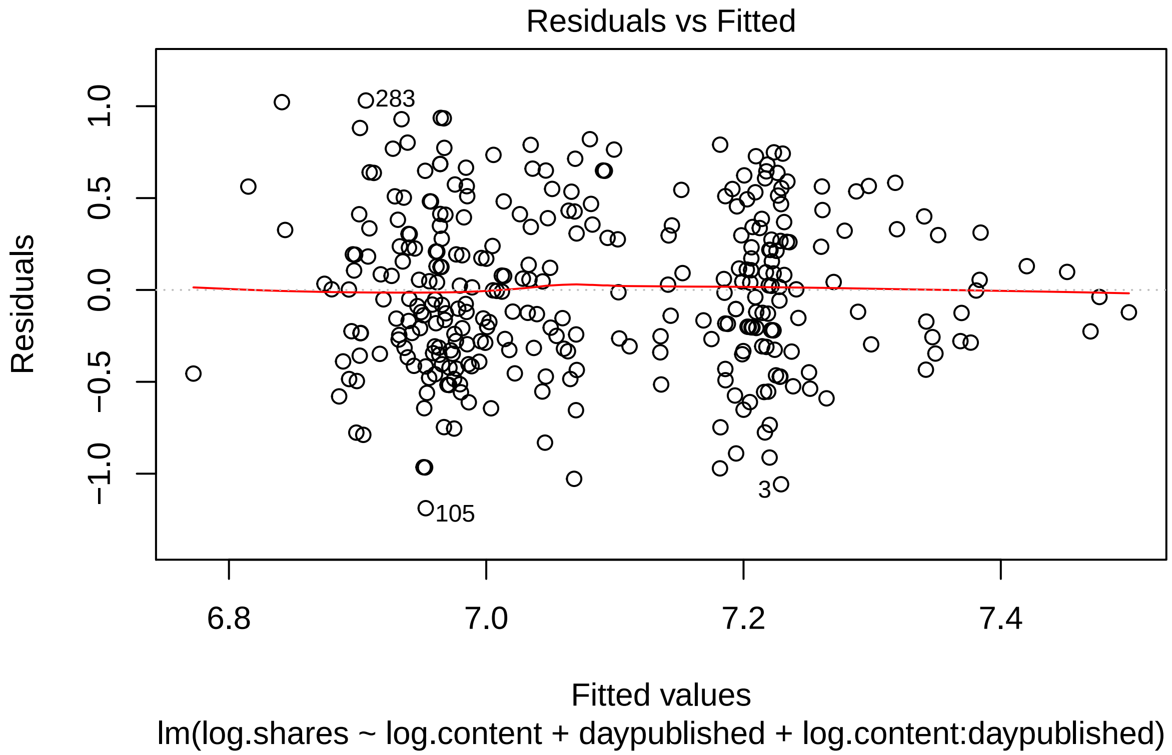
```
summary(interact.mod)
```

```
##
## Call:
## lm(formula = log.shares ~ log.content + daypublished + log.content:daypublished,
##     data = social.transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18776 -0.31318 -0.01091  0.32539  1.03094
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          8.7989     0.8041  10.943   <2e-16 ***
## log.content                         -0.2757     0.1308  -2.109   0.0358 *
## daypublishedMonday                  -1.4181     0.9209  -1.540   0.1246
## daypublishedSaturday                -0.7867     1.2069  -0.652   0.5150
## daypublishedSunday                  -3.0013     1.7221  -1.743   0.0824 .
## daypublishedThursday                -1.6964     0.9594  -1.768   0.0780 .
## daypublishedTuesday                 -1.1857     0.9819  -1.208   0.2281
## daypublishedWednesday               -2.4150     0.9697  -2.490   0.0133 *
## log.content:daypublishedMonday       0.2490     0.1492   1.669   0.0962 .
## log.content:daypublishedSaturday    0.1717     0.1930   0.890   0.3744
## log.content:daypublishedSunday      0.4799     0.2690   1.784   0.0754 .
## log.content:daypublishedThursday    0.2533     0.1550   1.635   0.1032
## log.content:daypublishedTuesday     0.1765     0.1583   1.115   0.2659
## log.content:daypublishedWednesday   0.3711     0.1560   2.379   0.0180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4468 on 308 degrees of freedom
## Multiple R-squared:  0.09644,    Adjusted R-squared:  0.0583
## F-statistic: 2.529 on 13 and 308 DF,  p-value: 0.002585
```

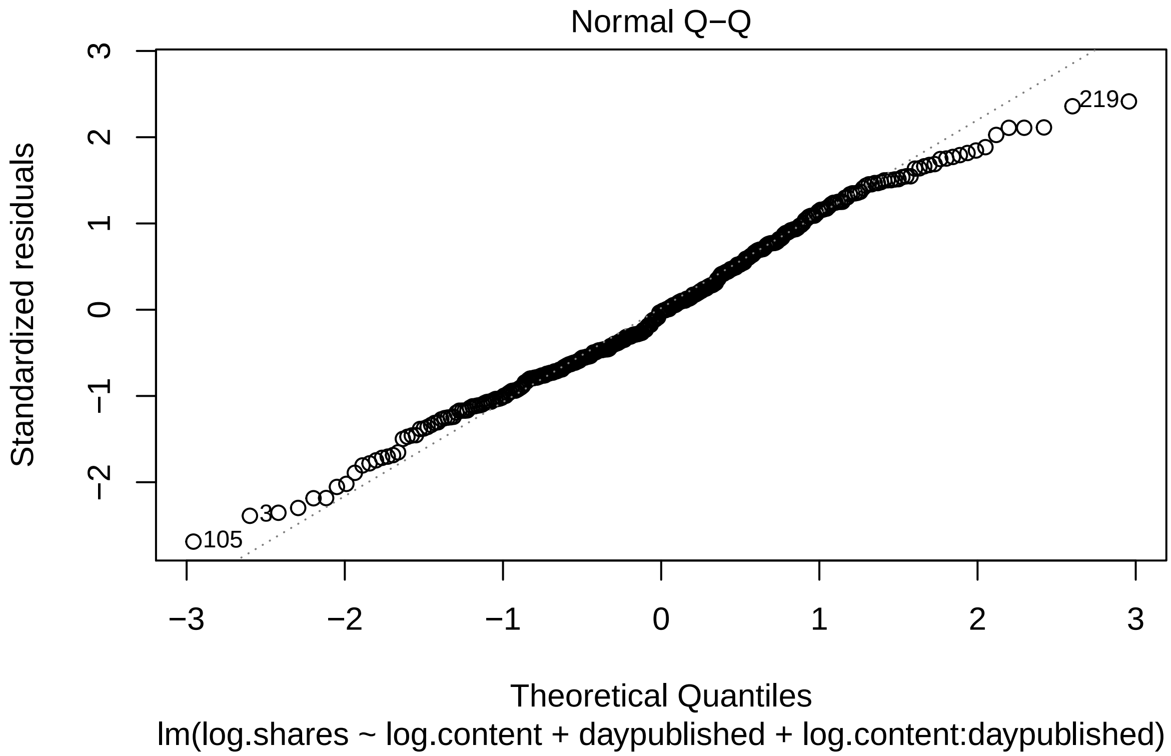
From the chart (non-parallel slopes) and numerical summary (significant interaction coefficients), one can notice significant interaction between the log(content) and daypublished variables. As such, the final model will be a linear regression model with interaction involving log(shares), log(content), and daypublished.

We then inspect the residual and qq plots of the final model as follows:

```
full.mod <- lm(log.shares ~ log.content + daypublished + log.content:daypublished,
               data = social.transform)
plot(full.mod,
     which=1)
```



```
plot(full.mod,
     which=2)
```



On the residual plot, we notice that apart from three marked outliers (rows 3, 105, and 283), the constant spread, independence, and mean zero assumptions are reasonably justified. On the QQ plot, we noticed that aside from slight deviation on the tail ends and outliers from rows 3, 105, and 219, the normaility assumption is reasonably justified as well. As such, the regression analysis summary from our final selected model appears as the follows:

```
summary(full.mod)
```

```
##
## Call:
## lm(formula = log.shares ~ log.content + daypublished + log.content:daypublished,
##     data = social.transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18776 -0.31318 -0.01091  0.32539  1.03094
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          8.7989     0.8041  10.943   <2e-16 ***
## log.content                         -0.2757     0.1308  -2.109   0.0358 *
## daypublishedMonday                  -1.4181     0.9209  -1.540   0.1246
## daypublishedSaturday                -0.7867     1.2069  -0.652   0.5150
## daypublishedSunday                  -3.0013     1.7221  -1.743   0.0824 .
## daypublishedThursday                -1.6964     0.9594  -1.768   0.0780 .
## daypublishedTuesday                 -1.1857     0.9819  -1.208   0.2281
## daypublishedWednesday               -2.4150     0.9697  -2.490   0.0133 *
## log.content:daypublishedMonday      0.2490     0.1492   1.669   0.0962 .
## log.content:daypublishedSaturday    0.1717     0.1930   0.890   0.3744
## log.content:daypublishedSunday      0.4799     0.2690   1.784   0.0754 .
## log.content:daypublishedThursday    0.2533     0.1550   1.635   0.1032
## log.content:daypublishedTuesday     0.1765     0.1583   1.115   0.2659
## log.content:daypublishedWednesday   0.3711     0.1560   2.379   0.0180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4468 on 308 degrees of freedom
## Multiple R-squared:  0.09644,    Adjusted R-squared:  0.0583
## F-statistic: 2.529 on 13 and 308 DF,  p-value: 0.002585
```

This appears to be a relatively reasonable model for predicting share numbers, as when compared to other models tried, this resulted in the highest R-squared value, most significant coefficients for all variables, and an F-statistic p-value of ~0.0026 < a = 0.05. Additionally, the transformed variables more reasonably satisfied assumptions when compared to non-transformed counterparts.

Prediction

With this signficiant model that reasonably satisfies all assumptions, the client is interested in predicting share numbers for a article that has 627 words, 3 images, and was published on a Saturday.

The predicted share count can be calculated with: `log(shares) = 8.799-0.276*log(627)-0.787*1+0.172*log(627)*1`

When computed, the resulting shares amount is approximately **1544 shares**

It should be noted that the value of '3 images' was omitted from the equation, as it is not a component of the model.

Discussion

Within this project, we looked to model a given article's share count when considering its word count, image count, and day published. While there is a relationship between share count and the previously mentioned explanatory variables, a series of transformations had to be performed to the quantitiative explanatory variables to create the best possible linear regression model that satisfied the highest number of assumptions and created a significant model. No issues of multicollinearity between explanatory variables was found, but the variable log(images) was omitted for the sake of simplicity and assumptions. A large potential limitation of this model lies in the transformation of many variables, which could influence the output of share amount and/or explanatory variables. The negative coefficient of log(content) could be an area of potential exploration as well. Additionally, potential missing data such as article topic is missing from the dataset. It may be valuable to know whether or not articles about pop culture will naturally have more shares than those about politics, for example.

In all, the tracking of trends on social media platforms continues to be a vital component in learning about human culture, beliefs, and values. In an effort to reach a wider audience, media sites should be aware of how the content and key components of every article influence its popularity.