Predicting Articles' Social Media Success

Dominic Hoar-Weiler dhoarwei

Due Wed, March 13, at 11:59PM

Contents

```r
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("interactions")
library("leaps")
```

Introduction

Social Media is a monolithic industry in the modern age. It is practically woven into the fabric of today's society. As such, it has become a useful tool for companies to market products and ideas. This is no different for news companies, who spread articles and news on social media to increase engagement. As such, it's useful to know what factors lead to increased engagement with the articles. For the purpose of this investigation, we will examine factors that help predict that number of shares an article receives.

Exploratory Data Analysis

DATA

In the social media dataset we are exploring, we analyze a random sample of 388 articles that were published on the site Mashable over the course of 2 years. Each article has four variables associated with it. Due to our interest in predicting the amount of shares an article receives, we examine the relationship between number of shares and three explanatory variables: content, images, and daypublished. We summarize the variables as follows:

Shares: number of shares the article has on social media

Content: number of words in the article

Images: number of images in the article

DayPublished: day of the week the article was originally published (Monday,. . . ,Sunday)

The first few lines of the data appear as follows:

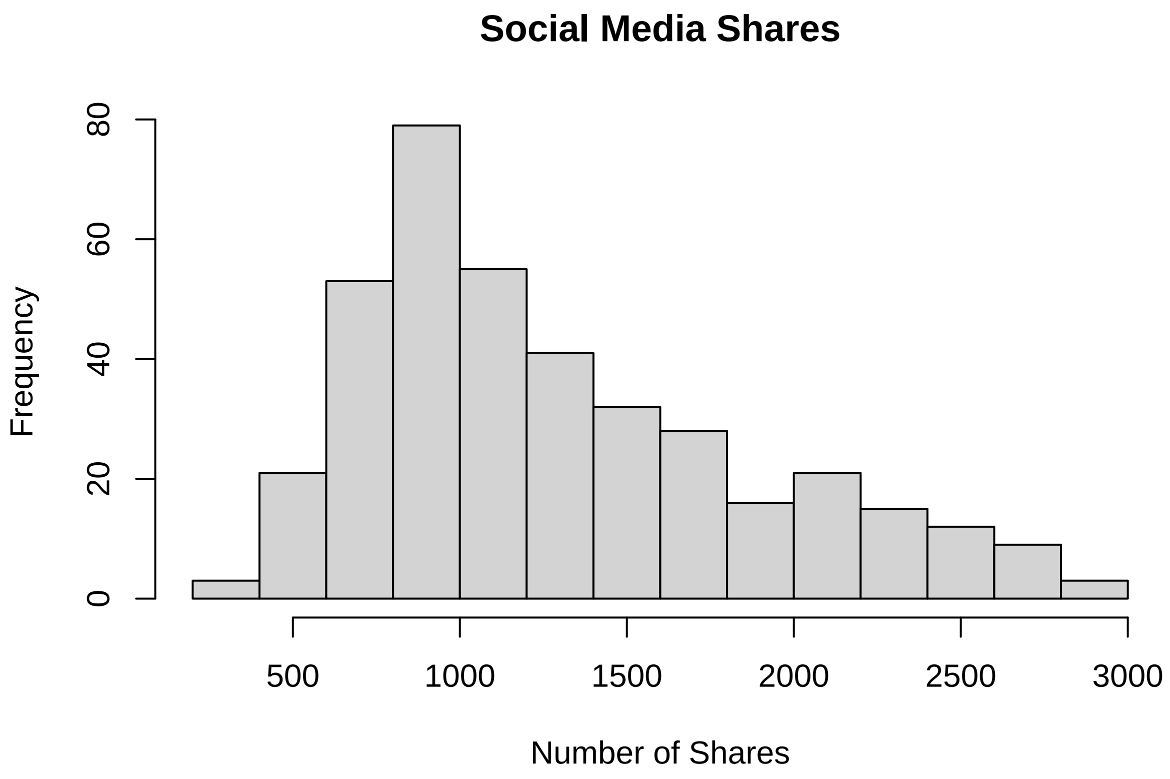
```
head(social)
```

```
## # 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
```

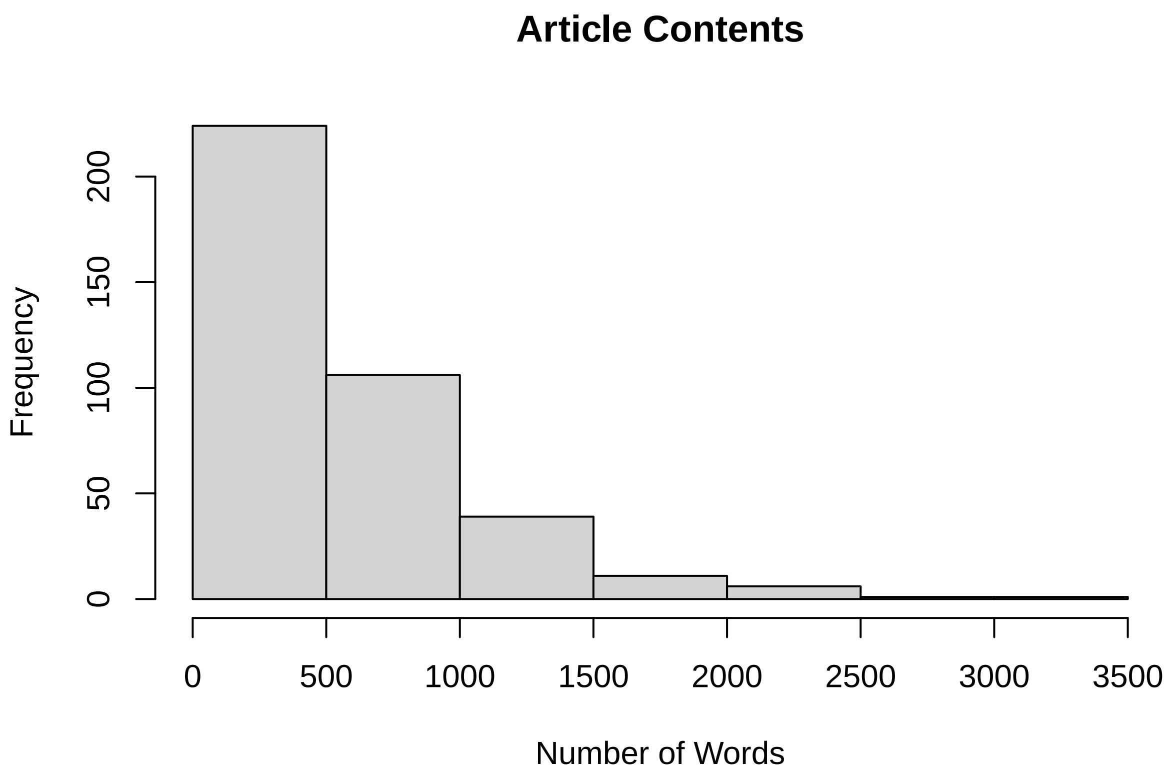
UNIVARIATE EXPLORATION

First, we will explore each variable individually. We'll use histograms to explore our quantitative variables and a barplot to explore our categorical variable.

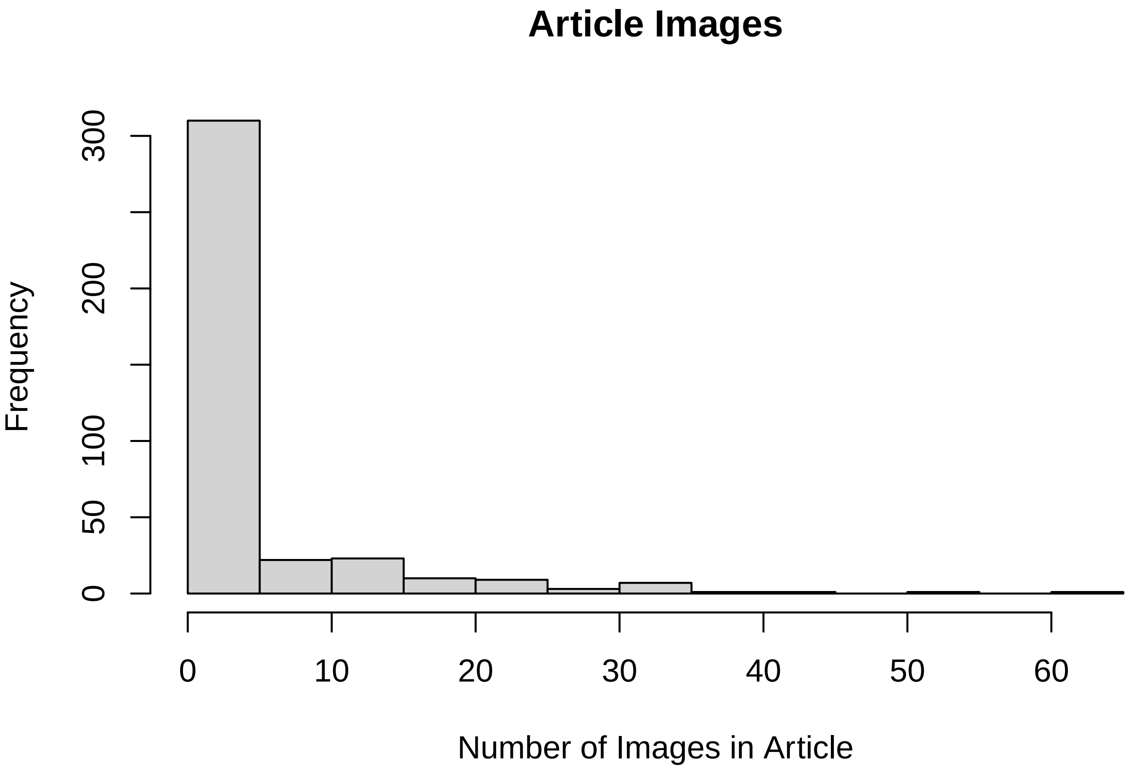
```
hist(social$shares,
     main = "Social Media Shares",
     xlab = "Number of Shares")
```



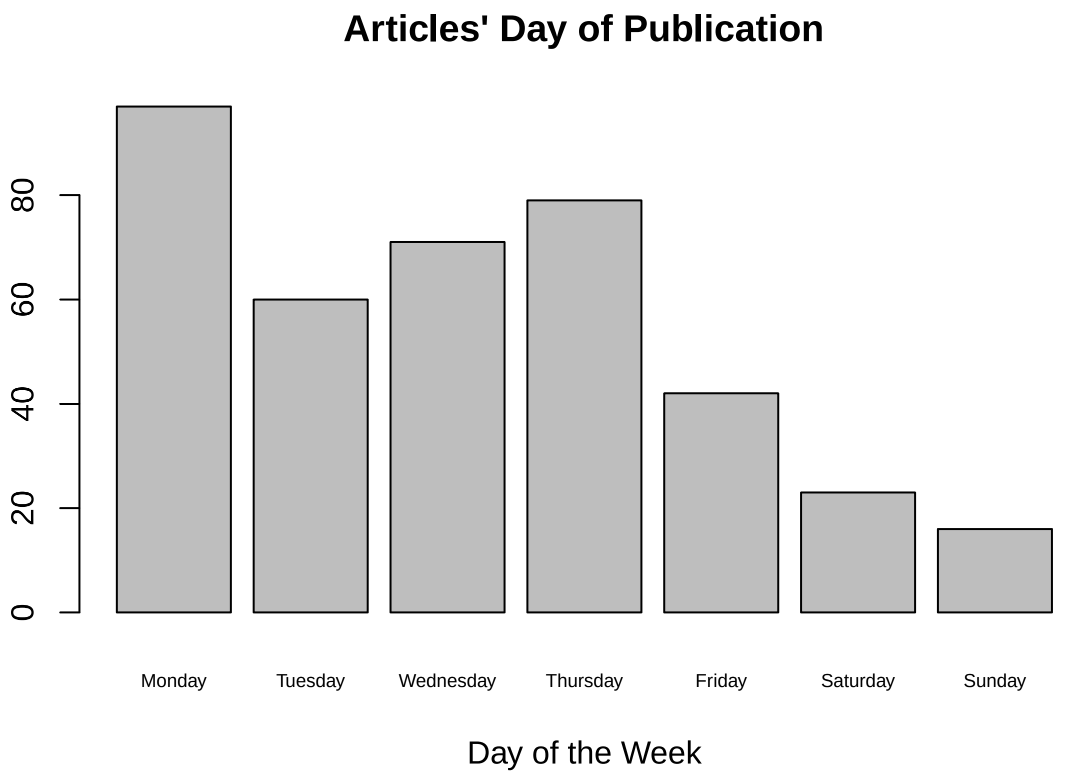
```r
hist(social$content,
     main = "Article Contents",
     xlab = "Number of Words")
```



```r
hist(social$images,
     main = "Article Images",
     xlab = "Number of Images in Article")
```



```r
social$sequenced_days <- factor(social$daypublished, levels=c("Monday","Tuesday","Wednesday","Thursday"
barplot(table(social$sequenced_days),
        main = "Articles' Day of Publication",
        xlab = "Day of the Week",
        cex.names = 0.6)
```



We supplement the univariate graphical summary with numerical summaries, as follows:

For Shares:

```
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:

```
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:

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

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

For DayPublished:

```
table(social$sequenced_days)
```

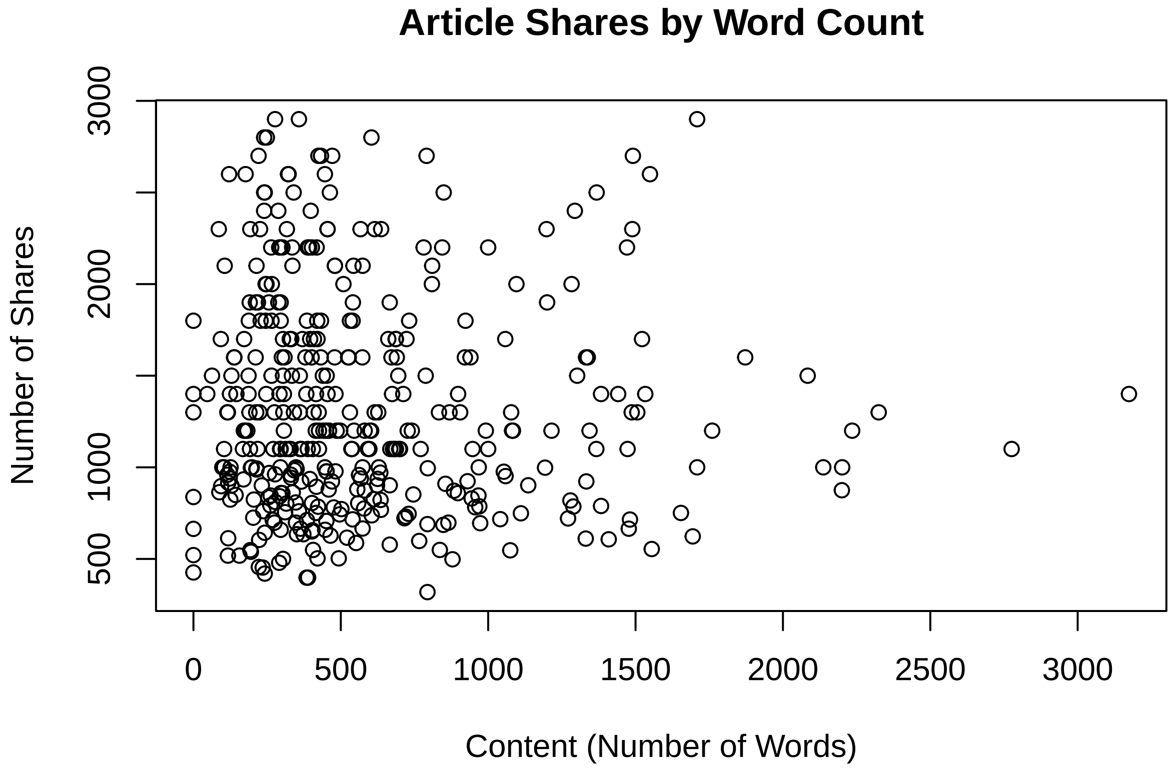
```
## 
##    Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday 
##        97        60        71        79        42        23        16
```

After exploring the graphs we made in addition to our summary statistics for each variable, we can make the following observations. The distribution of article **shares** is unimodal and right-skewed. The distribution is centered around 1200 shares, and the average amount of shares is 1325.1. There are also no clear outliers in the distribution. The distribution of the articles' **content**, which is measured in word count, is also unimodal and right-skewed, with no clear outliers. The distribution is centered around 433 words, and the average word count of the articles is 586.1 words. The distribution of **images** is unimodal and heavily right-skewed. The vast majority of articles have less than five images. There are some outliers in this distribution. Namely, there is an article with 51 images and another one with 61 images. That being said, the distribution is centered around 1 image and the average amount of images in the articles is 4.433. Now, when looking at the **days of the week** that articles were published, we see that more articles were published towards the start of the week (Monday,Tuesday. . .) than the end of the week (. . . Saturday,Sunday). For example, 97 articles were published on Monday and 60 articles were published on Tuesday while only 23 articles were published on Saturday and 16 articles were published on Sunday.

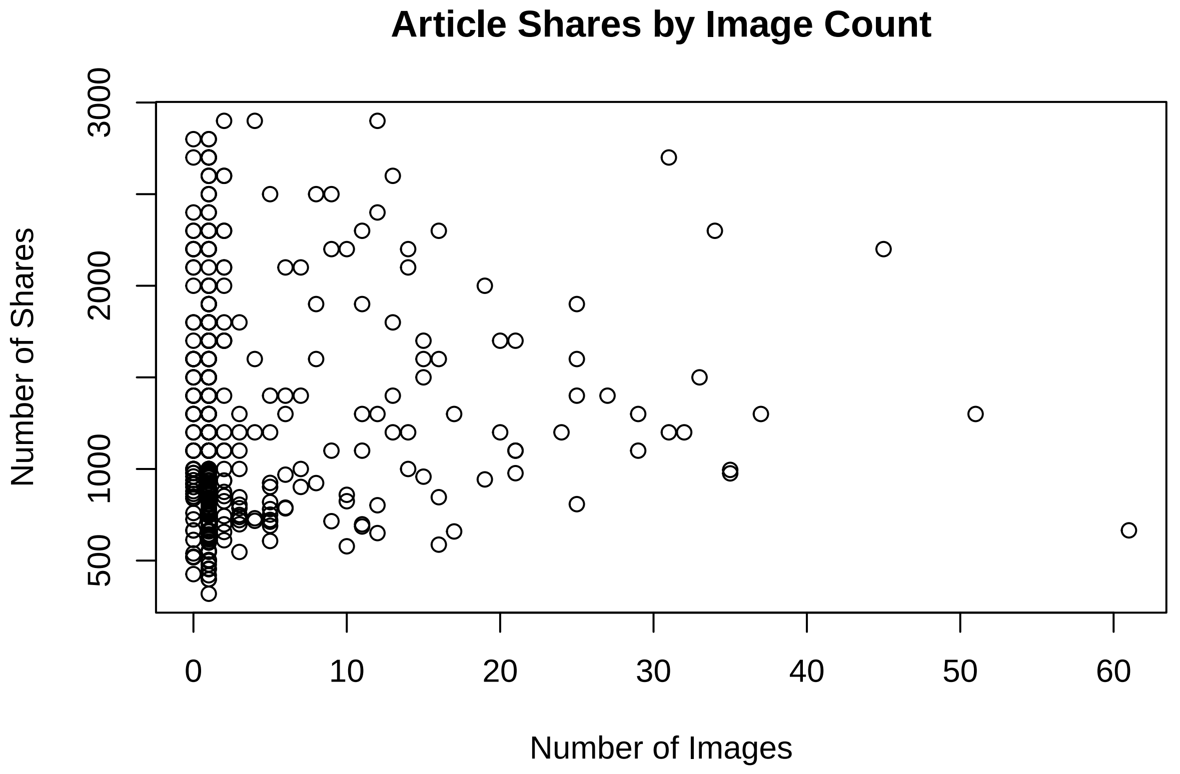
BIVARIATE EXPLORATION

Now, we'll explore how each of the predictor variables relates to the number of shares of the articles.

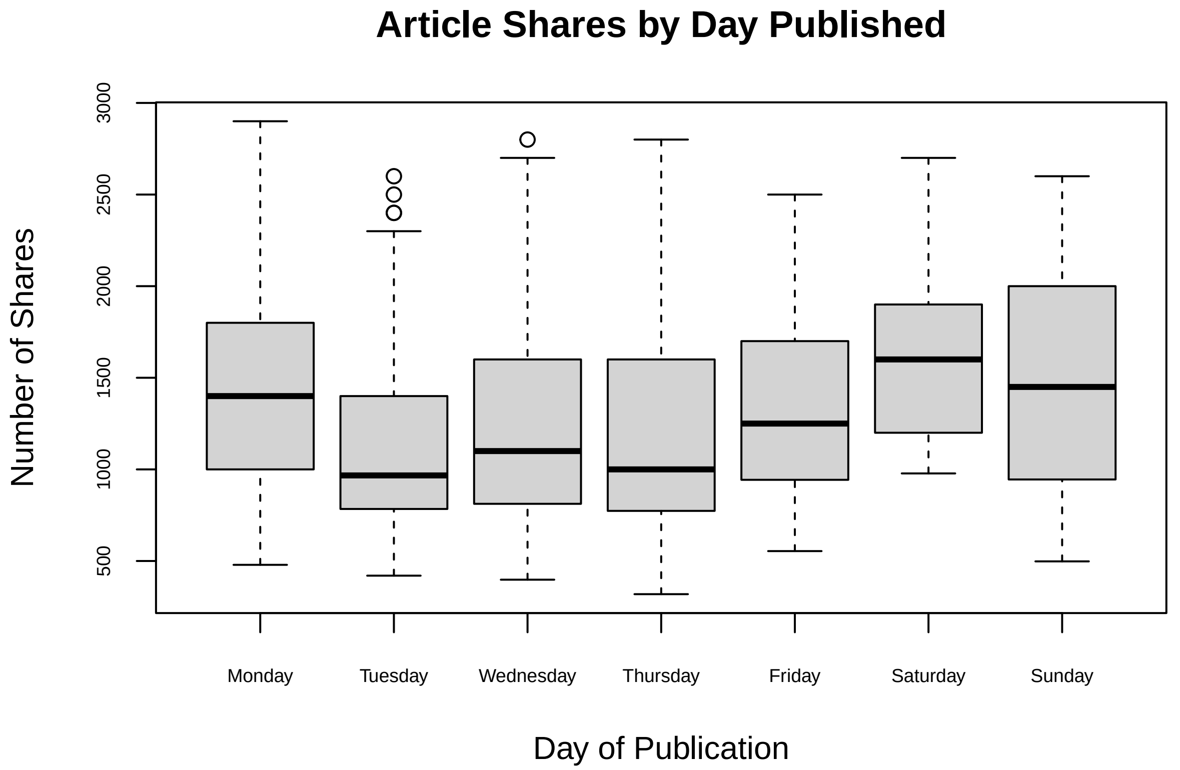
```
plot(shares ~ content,data=social,main="Article Shares by Word Count",
            xlab="Content (Number of Words)",
            ylab="Number of Shares")
```



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



```
boxplot(shares ~ sequenced_days,data=social,main="Article Shares by Day Published", xlab="Day of Publica
        ylab="Number of Shares",
        cex.axis=0.6)
```



Correlation Matrix for the Quantitative Variables:

```
social.quant <- subset(social,select = c(shares,content,images))
cor(social.quant)
```

```
##              shares     content     images
## shares   1.00000000 -0.02593102 0.04562069
## content -0.02593102  1.00000000 0.37423401
## images   0.04562069  0.37423401 1.00000000
```

By analyzing the graphs above, we can see the following observations. First, there seems to be no linear relationship or relationship at all between the **word count** (content) of the articles and the number of shares it receives. Graphically, there is no clear relationship between the two variables. The correlation matrix supports this, as the correlation coefficient between content and shares is -0.026. This suggests that there might be a very slight negative relationship. However, this number is too close to zero to suggest a statistically significant relationship.

Second, there seems to be no significant relationship between **number of images** in the article and the number of shares it receives either. By looking at the graph, there looks like there could be a slight positive relationship, suggesting that more images would lead to more shares. However, by looking at the correlation matrix, we see that these variables have a correlation coefficient of 0.046, which is too close to zero to suggest any sort of relationship.

Lastly, there does seem to be a slight correlation between **day of publication** and number of shares. Articles published on Mondays, Saturdays and Sundays all seem to have slightly higher number of shares. However, if there is a correlation, it is very slight as all the Q1/Q3 boxes in the plot overlap.

Modeling

After exploring the relationships among our variables, we look to build a linear regression model to predict the number of shares an article receives. We saw in our bivariate exploratory data analysis that content and images have little to no statistically significant relationship with the number of shares an article receives. However, we did find that the day an article was published had a correlation with the number of shares it received. As such, I'm considering dropping content and images as predictor variables in the model. To confirm their lack of significance, I conducted a multiple linear regression that predicts shares from content, images and daypublished.

Before I make this model, I want to make sure that there is no multicollinearity problem between content and images as they had a correlation coefficient of 0.374. Below are the GVIF's of the model.

```
shares.fromall <- lm(shares ~ content + images + daypublished,data=social)
car::vif(shares.fromall)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## content      1.172911  1        1.083010
## images       1.187833  1        1.089878
## daypublished 1.031635  6        1.002599
```

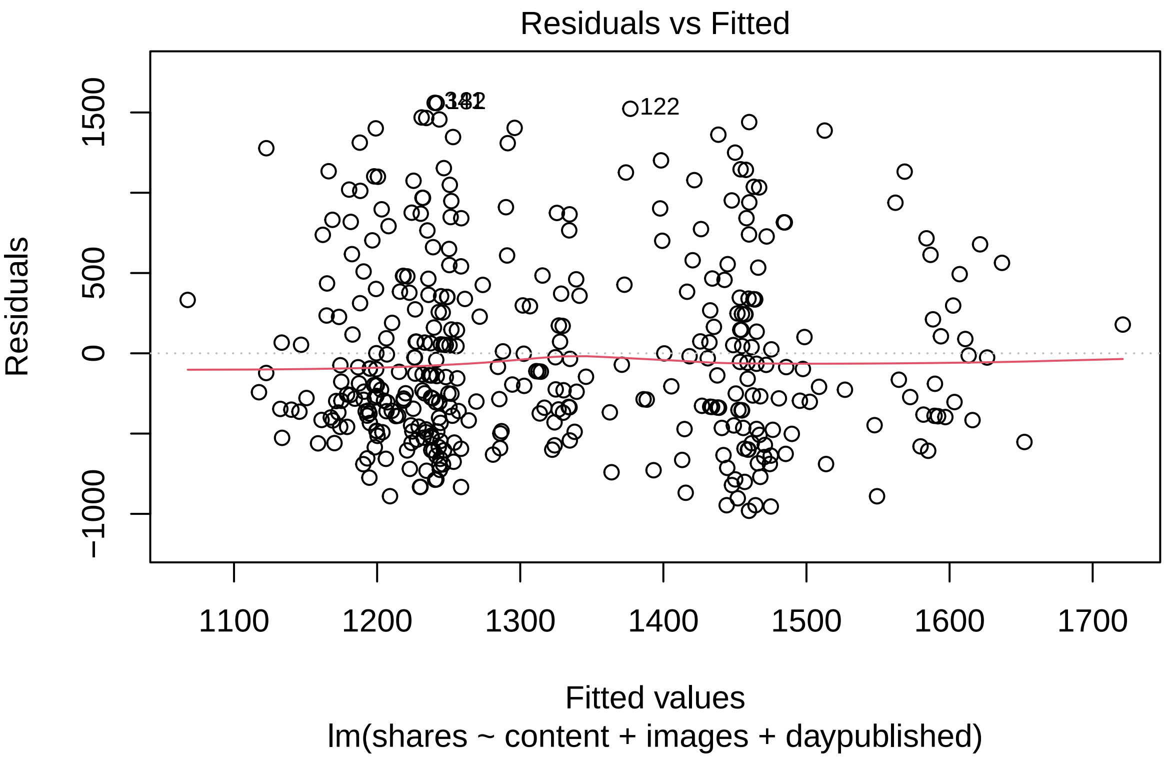
Since none of the GVIF's are above 2.5, we do not have a multicollinearity problem, and we can proceed with the model. Below is the summary of the model.

```
summary(shares.fromall)
```

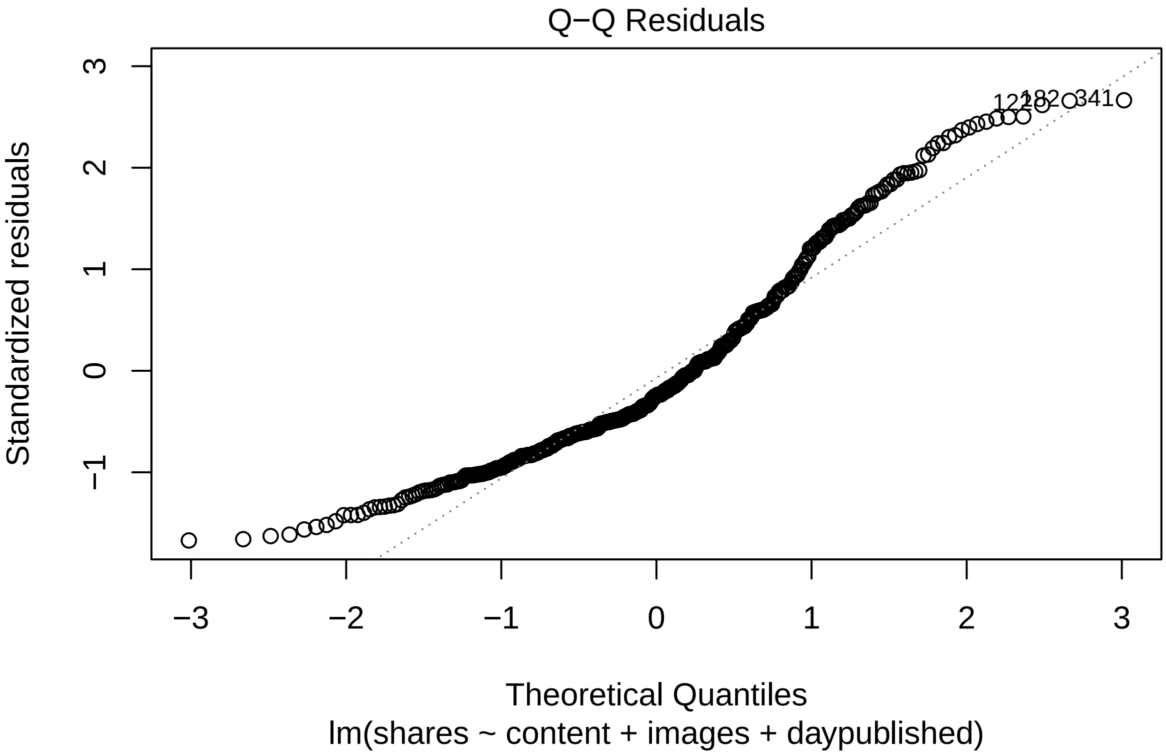
```
## 
## Call:
## lm(formula = shares ~ content + images + daypublished, data = social)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -980.9 -421.0 -143.8  347.6 1559.8 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1347.80387   96.84930  13.917   <2e-16 ***
## content               -0.06846    0.06905  -0.992    0.322    
## images                 4.71893    3.98135   1.185    0.237    
## daypublishedMonday   127.26428  109.32710   1.164    0.245    
## daypublishedSaturday 270.25665  153.43482   1.761    0.079 .  
## daypublishedSunday   151.89809  173.86151   0.874    0.383    
## daypublishedThursday -89.17492  113.20046  -0.788    0.431    
## daypublishedTuesday -141.35899  118.84820  -1.189    0.235    
## daypublishedWednesday -95.91981  115.25370  -0.832    0.406    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 590.1 on 379 degrees of freedom
## Multiple R-squared:  0.04818,    Adjusted R-squared:  0.02809 
## F-statistic: 2.398 on 8 and 379 DF,  p-value: 0.01562
```

Once again, we see that daypublished is the only significant predictor of the number of shares an article receives.

```
plot(shares.fromall,which=1)
```

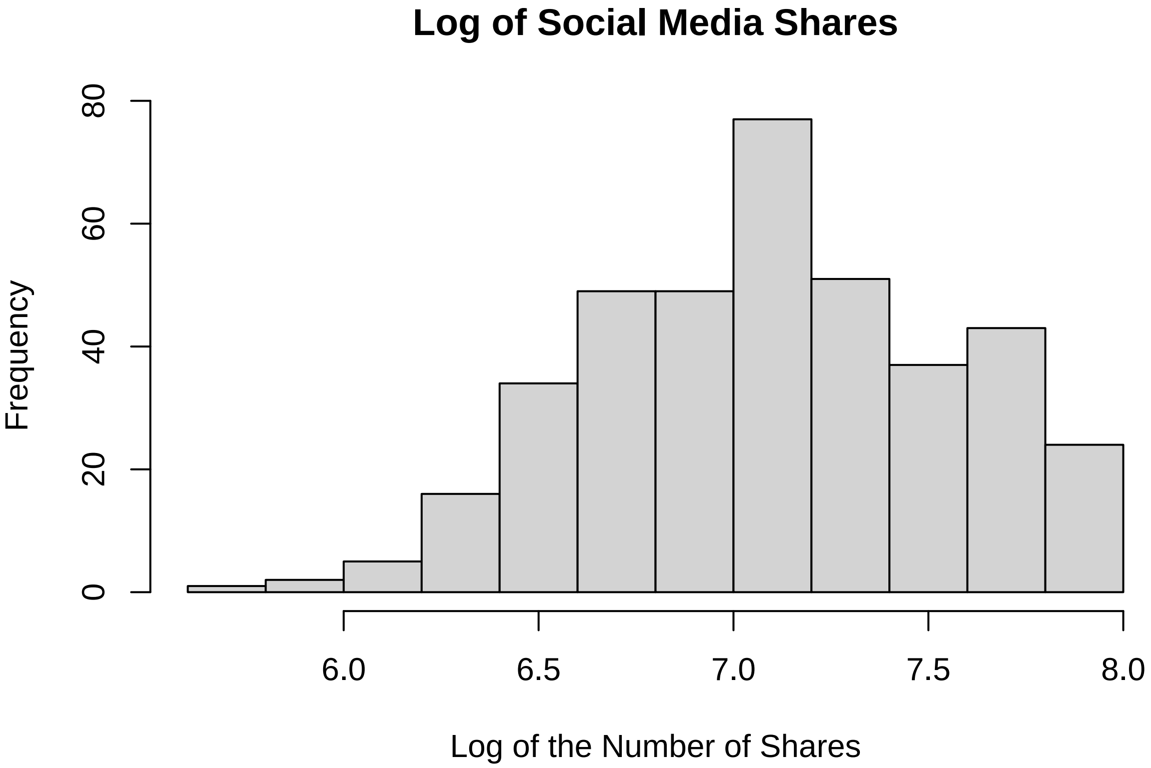


```r
plot(shares.fromall,which=2)
```



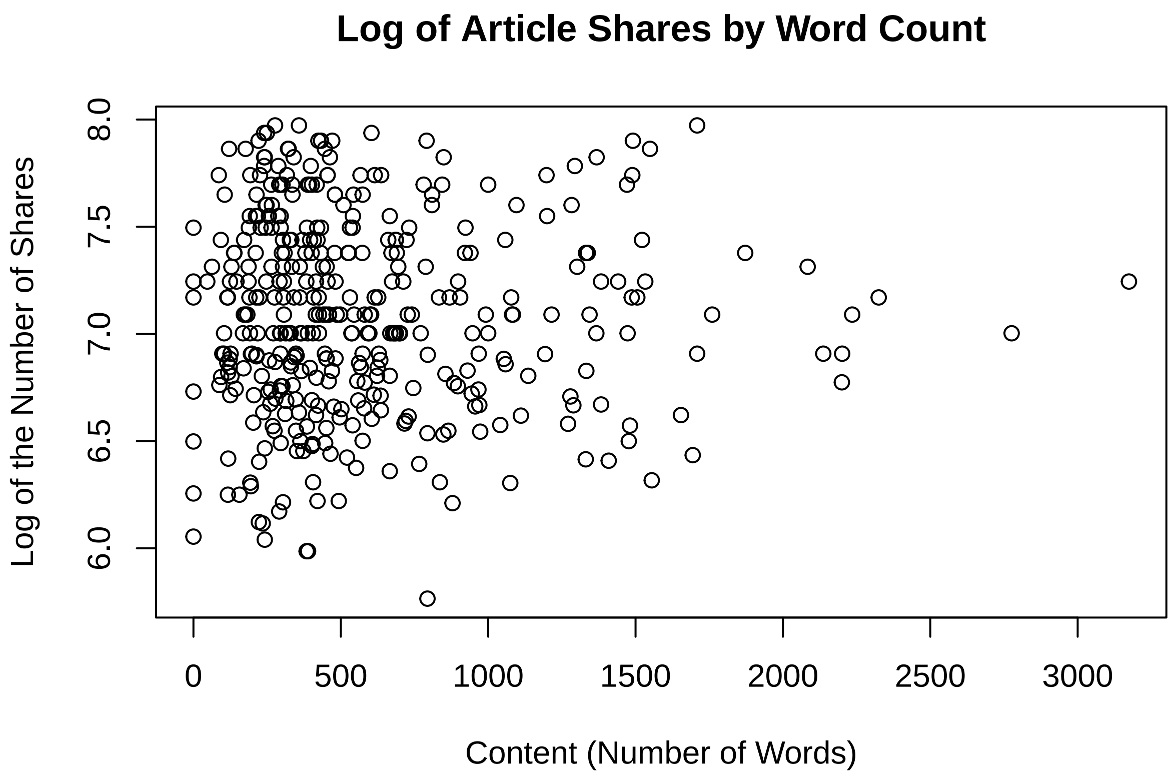
After looking at the residual plot, we observe the following. First, the residuals for this model seem to meet the constant spread, independence and mean zero assumptions. However, looking at the QQplot, we see deviations at the end which may invalidate the normality condition. As such, to improve diagnostics and significance, I've decided to transform the right-skewed variable shares, using a log transformation. Below is the distribution of log(shares).

```
social$log.shares <- log(social$shares)
hist(social$log.shares,
     main = "Log of Social Media Shares",
     xlab = "Log of the Number of Shares")
```

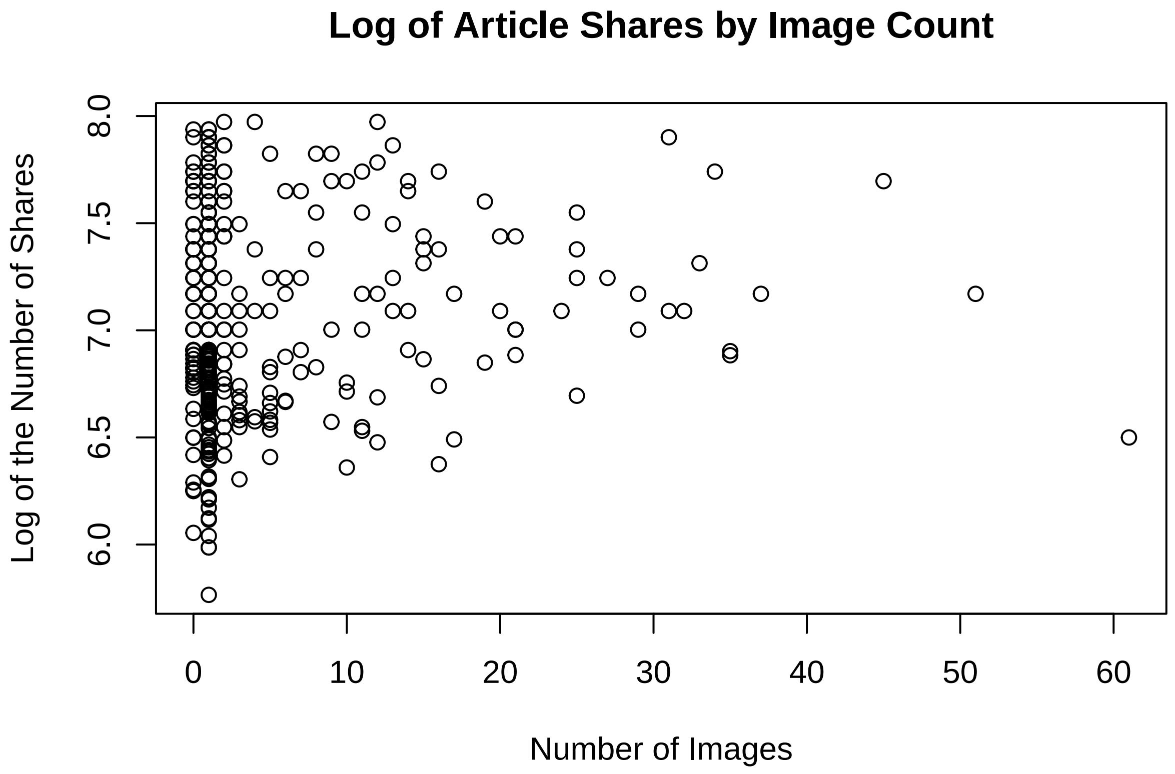


The distribution of the log of shares is unimodal and very slightly left-skewed, with no significant outliers. However, this distribution is far more symmetrical than the untransformed variable shares. As such, we proceed to bivariate EDA on log.shares.

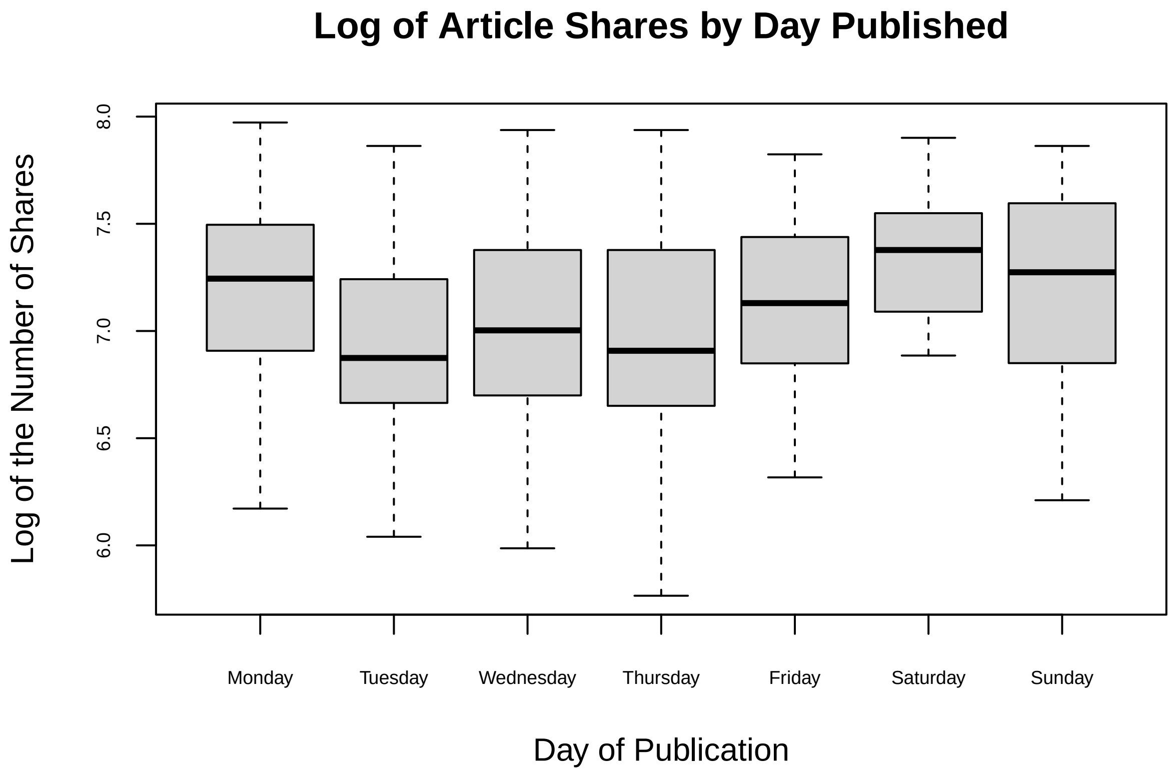
```
plot(log.shares ~ content,data=social,main="Log of Article Shares by Word Count",
            xlab="Content (Number of Words)",
            ylab="Log of the Number of Shares")
```



```
plot(log.shares ~ images,data=social,main="Log of Article Shares by Image Count",
            xlab="Number of Images",
            ylab="Log of the Number of Shares")
```



```
boxplot(log.shares ~ sequenced_days,data=social,main="Log of Article Shares by Day Published", xlab="Da
        ylab="Log of the Number of Shares",
        cex.axis=0.6)
```



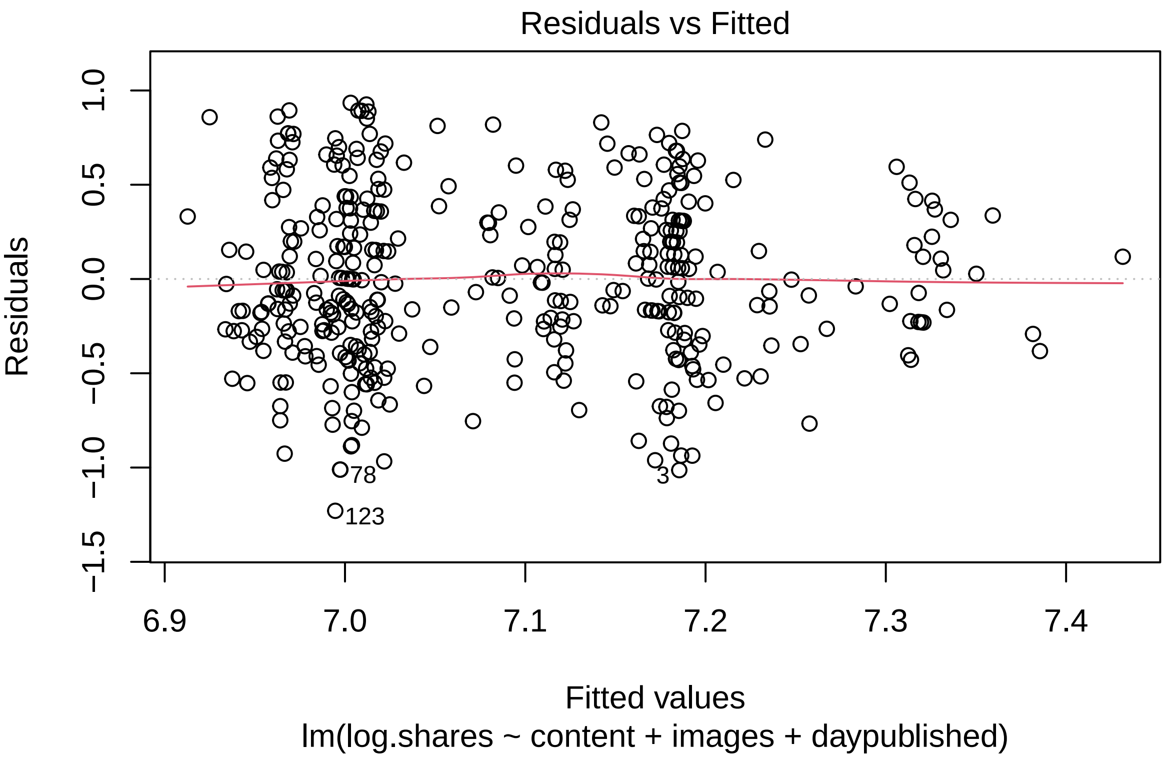
```
social.quantlog <- subset(social,select = c(log.shares,content,images))
cor(social.quantlog)
```

```
##               log.shares      content     images
## log.shares   1.000000000 -0.009587799 0.06285946
## content     -0.009587799  1.000000000 0.37423401
## images       0.062859462  0.374234014 1.00000000
```

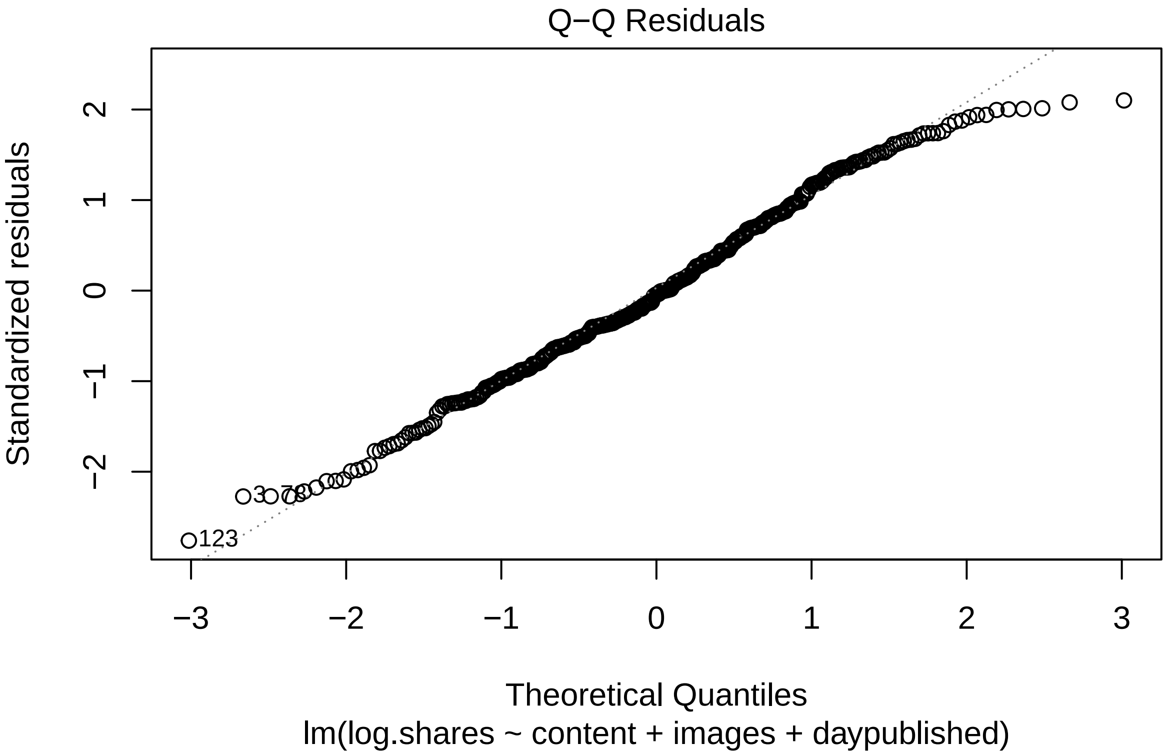
After analyzing the graphs and the correlation matrix involving the bivariate analysis of log(shares), we make the following observations. First, there seems to be no significant correlation between content and the log of shares (a correlation coefficient of -0.0096). Similarly, there seems to be no significant correlation between images and the log of shares (a correlation coefficient of 0.0629). However, we do still see a slight relationship between day of publication and the log of shares, with Monday, Saturday and Sunday having slightly higher log of share counts.

Before formally deciding to remove content and images as predictors in the model, we conduct another linear regression model, predicting the log of shares from content, images and daypublished.

```
logshares.fromall <- lm(log.shares ~ content + images + daypublished, data = social)
plot(logshares.fromall,which=1)
```



```
plot(logshares.fromall,which=2)
```



Upon inspecting the residual plot, we see that the independence, constant spread and mean zero assumptions hold. This is because the residual plot has fairly equal spread across the graph, seems relatively patternless and is centered at a residual of 0. The QQplot with the transformed log(shares) model seems far more likely to validate the normality assumption, as almost all the points are centered on the line. Below is the summary for the linear regression model.

```
summary(logshares.fromall)
```

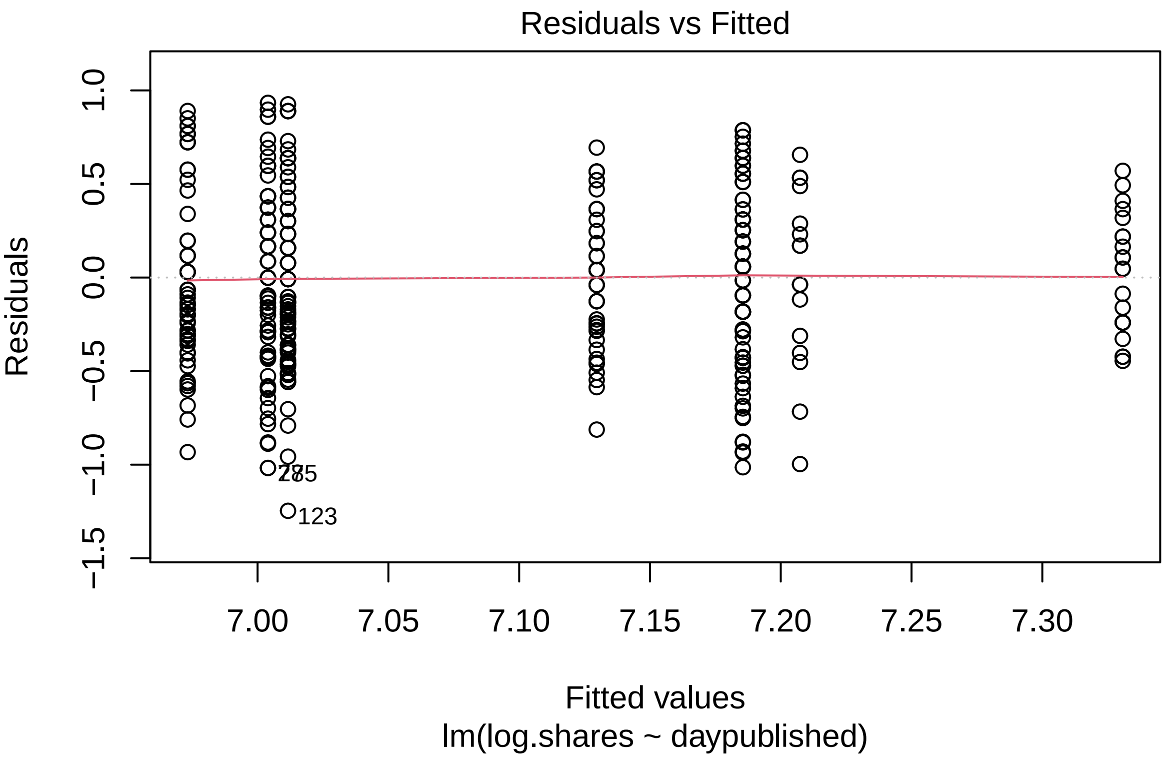
```
##
## Call:
## lm(formula = log.shares ~ content + images + daypublished, data = social)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22940 -0.29000 -0.01905  0.32063  0.93423
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.128e+00  7.358e-02  96.877   <2e-16 ***
## content              -3.957e-05  5.246e-05  -0.754   0.4511
## images                4.284e-03  3.025e-03   1.416   0.1576
## daypublishedMonday    6.440e-02  8.306e-02   0.775   0.4386
## daypublishedSaturday  2.048e-01  1.166e-01   1.757   0.0798 .
## daypublishedSunday    7.427e-02  1.321e-01   0.562   0.5743
## daypublishedThursday -1.065e-01  8.600e-02  -1.239   0.2162
## daypublishedTuesday  -1.564e-01  9.029e-02  -1.733   0.0840 .
## daypublishedWednesday -1.199e-01  8.756e-02  -1.369   0.1717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4484 on 379 degrees of freedom
## Multiple R-squared:  0.05864,    Adjusted R-squared:  0.03877
## F-statistic: 2.951 on 8 and 379 DF,  p-value: 0.003261
```

Once again, we observe that the beta estimates for content and images are not statistically significant, with p-values of 0.45 and 0.16 respectively. As such, it's now justifiable that we drop these predictors from our model. That being said, transforming shares to log(shares) helped increase the statistical significance of the daypublished predictor variable. We see Saturday and Tuesday's coefficients as being statistically significant at the 10% level. Additionally, the adjusted R-squared of this model is 0.03877 while the previous model's was 0.02809.

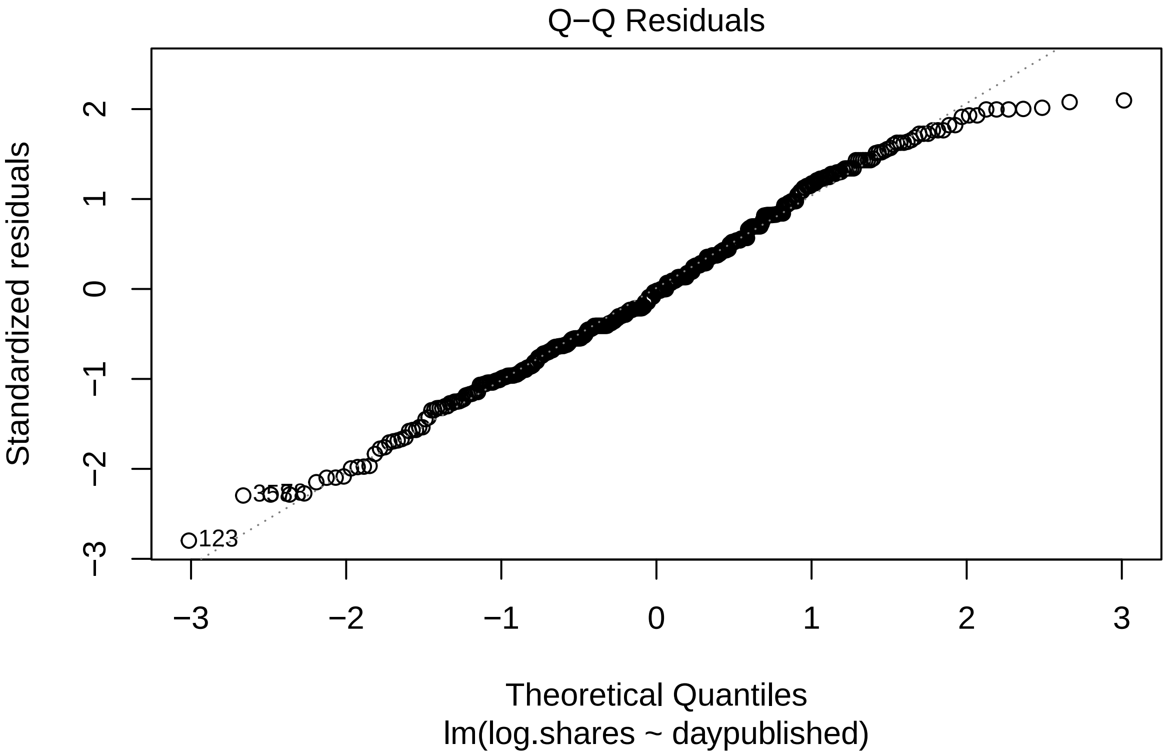
I also tried models with transformations for log(content) and log(images), but the models yielded similar results as the example above, so I've omitted them from the report. However, once again, they showed justification for the insignificance of content and images in predicting shares or log(shares).

Now, we conduct a linear regression model, dropping content and images as predictor variables. Below are the residual diagnostic plots for the regression model that predicts the log of shares from the day the article was published.

```
logshares.fromday.mod <- lm(log.shares ~ daypublished,data=social)
plot(logshares.fromday.mod,which=1)
```



```
plot(logshares.fromday.mod,which=2)
```



First, we will investigate the residual diagnostic plots from our model with DayPublished.

On the residual plot, we observe the constant spread, independence and mean zero assumptions are reasonably justified. While there is slight variation in the spreads, the spread of the residuals is consistent enough that we can justify this assumption. Additionally, there is no clear relationship between the residuals, justifying the independence assumption. Lastly, the mean is centered on the 0 line, justifying the mean assumption. Looking at the qqplot, we see that the normality assumption is justified as the majority of data points fall on the line.

Since all four of the error assumptions were justified, we proceeded with our model. Once again, below is the regression analysis summary of our chosen model.

```
summary(logshares.fromday.mod)
```

```
##
## Call:
## lm(formula = log.shares ~ daypublished, data = social)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24645 -0.30277 -0.01199  0.31225  0.93341
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.12965    0.06919 103.048   <2e-16 ***
## daypublishedMonday     0.05587    0.08282   0.675   0.5004
## daypublishedSaturday   0.20111    0.11631   1.729   0.0846 .
## daypublishedSunday     0.07772    0.13173   0.590   0.5556
## daypublishedThursday  -0.11801    0.08563  -1.378   0.1689
## daypublishedTuesday   -0.15638    0.09021  -1.734   0.0838 .
## daypublishedWednesday -0.12569    0.08729  -1.440   0.1507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4484 on 381 degrees of freedom
## Multiple R-squared:  0.05351,    Adjusted R-squared:  0.0386
## F-statistic:  3.59 on 6 and 381 DF,  p-value: 0.001788
```

We see that this model is significant because the regression F-test yielded a p-value of 0.001788, which means the model is significant at an alpha level of 0.01. Similarly, note that at least one of the beta coefficients is significant (Saturday, Tuesday) at the 0.1 alpha level, which means day published is a statistically significant predictor. Also note that the R^2 value of the model is 0.05351, which means that 5.35% of the variation in the log of the number of shares an article receives can be explained by its linear relationship with the day it was published. This is very small though, so it should be mentioned that this model is still a pretty weak predictor.

We also see that the positive coefficients for daypublishedMonday, daypublishedSaturday, and daypublishedSunday confirm our EDA results, which found slightly higher log of the number of shares for articles published on those days in comparison with Friday (which is used as the baseline in the model).

Prediction

Now that we have a model that reasonably satisfies all error assumptions, we are interested in predicting the number of shares that an article with 627 words, 3 images that was published on Saturday will receive.

The predicted log of shares is computed as follows:

```
7.12965 + 0*0.05587 + 1*0.20111 + 0*0.07772 + 0*-0.11801 + 0*-0.15638 + 0*-0.12569
```

```
## [1] 7.33076
```

To compute the number of shares this model yields, we raise e to this power:

```
exp(7.33076)
```

```
## [1] 1526.542
```

The predicted number of shares the article with 627 words and 3 images that was published on Saturday will receive is about 1527 shares. Note that since our model only used daypublished to predict the number of shares, the number of words and images in the article did not matter.

Discussion

Overall, we found that the number of shares an article receives is loosely correlated with the day on which it was published. Articles published on the weekend, Monday and Friday had slightly higher number of shares than those published during the middle of the week. We also found that there was no significant relationship between the number of shares an article received and the number of words or images it had.

Specifically, our model predicts the log of the number of shares an article receives from the day it was published. Transforming the response variable helped enhance the diagnostics of the plot; however, this does limit our predictions to the log of the number of shares an article receives.

Our model was also not very significant in predicting the number of shares, due to a small Rˆ2 value. Only 5.35% of the variation in the log of the number of shares could be explained by its linear relationship with the day it was published. In the future, exploring other pieces of data, such as the time of day the article was published and what topic the article was about, might help us yield a model that can help explain more of the variation in shares.