Predicting Online Article Shares by Content, Images, and Day Published

Steffi Chern
steffic

Contents

Introduction

Due to the technology innovations in the past decades, there is a notable rise in the number of users who tend to consume news from the internet or social networks. Knowing what affects users to share contents or not is crucial for anyone who seek to increase their content exposure on the internet. Different articles get shared by users in different amount of times, thus indicating that there may be some features about these articles that explain why users share some contents more times than the other ones. In this paper, we look into the potential features about the articles that may alter a user's decision of sharing it or not.

Exploratory Data Analysis

Data

In this research, we collect our sample of various online articles from Mashable. We examine the social media dataset that consists of a random sample of 388 online articles, 3 predictors (content, images, and day published), and 1 response variable (shares). To understand the potential relationships between the number of shares and the three predictors, we first summarize the variables, as below:

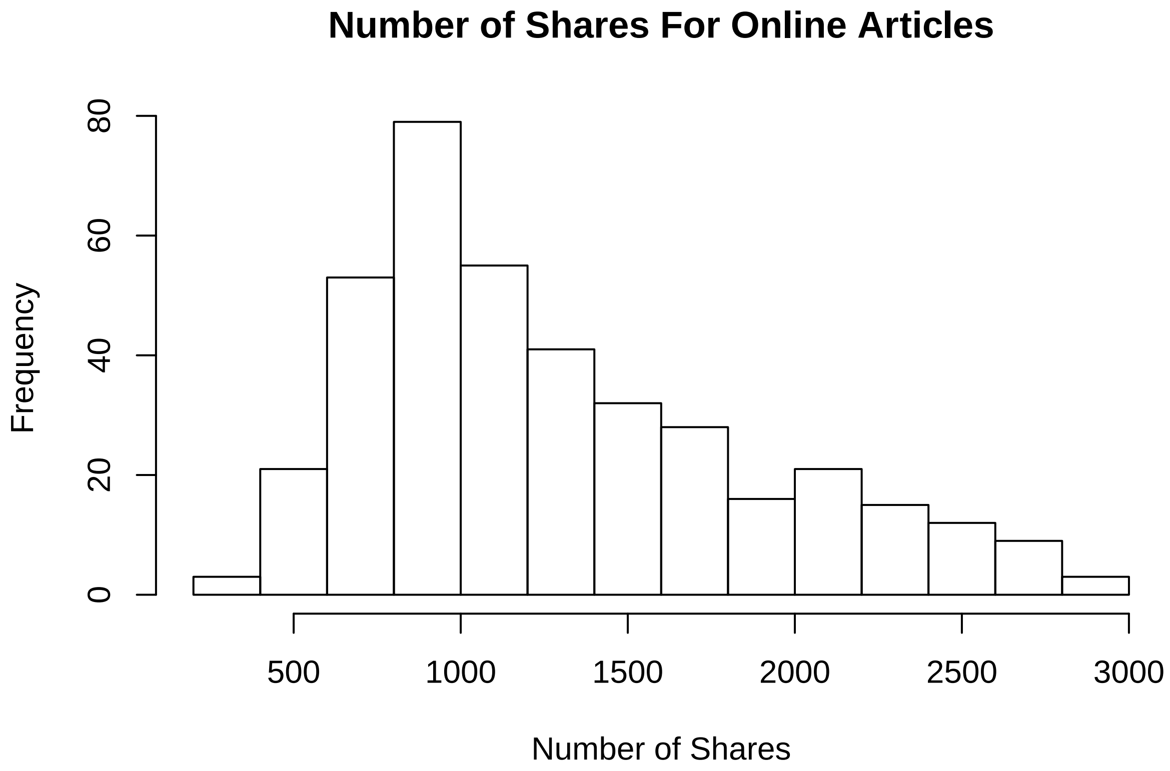
- Shares: number of times the online article was shared on social media
- Content: number of words in the online article
- Images: number of images in the online article
- Day Published: which day of the week the online article was published (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)

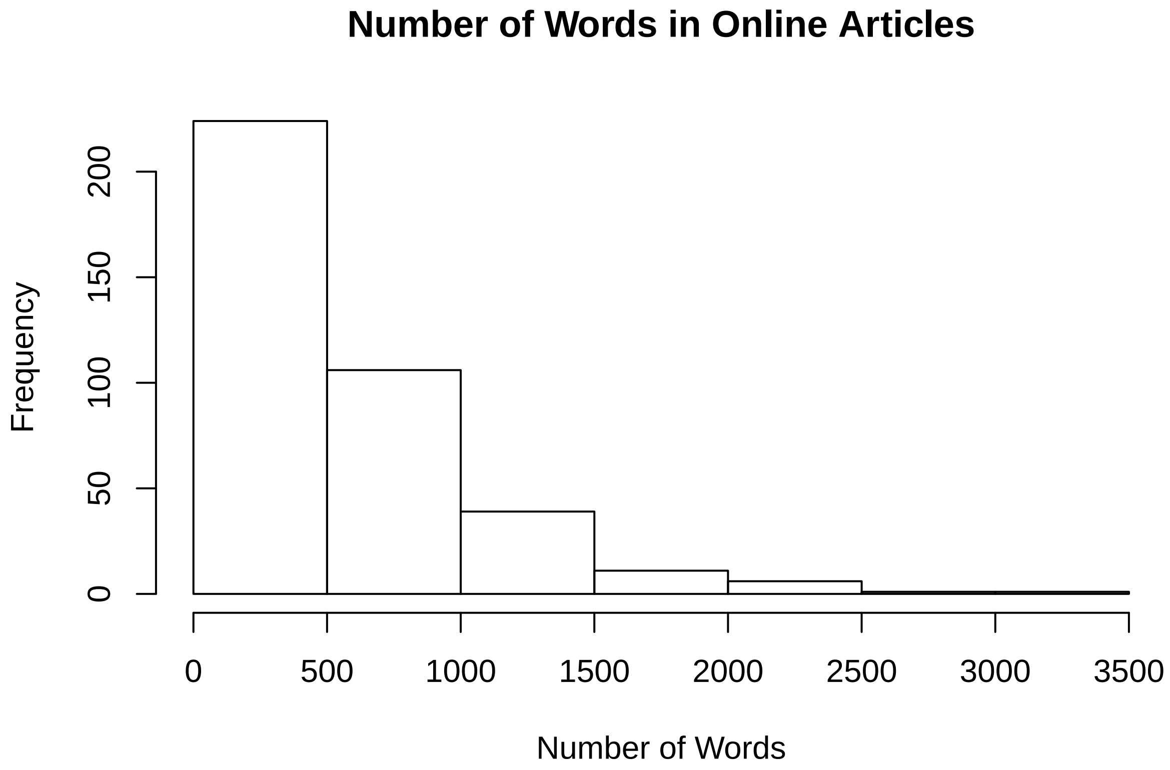
The first 10 data from the social media dataset:

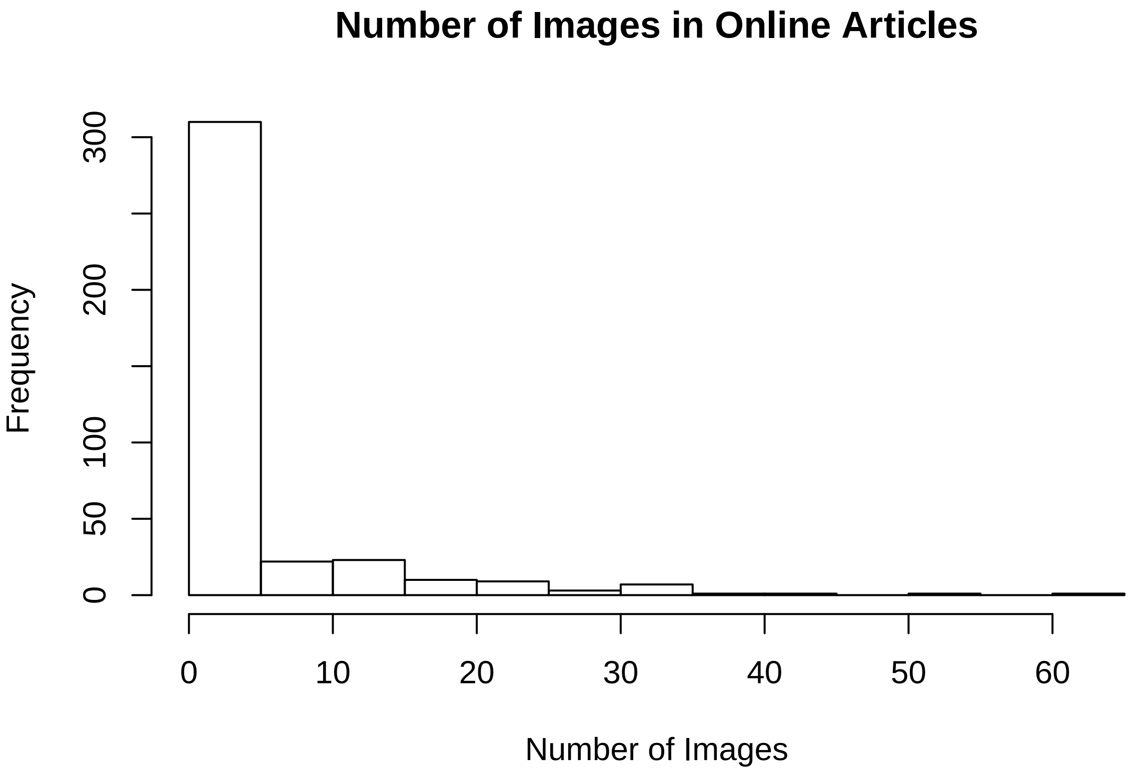
```
## # A tibble: 10 x 4
##    shares content images daypublished
##     <dbl>   <dbl>  <dbl> <chr>
##  1   1100     367      1 Monday
##  2   1400     712      1 Monday
##  3    479     291      1 Monday
##  4   2500     463      5 Monday
##  5   1200     498     13 Monday
##  6   1200    1084      1 Monday
##  7   1500     361      1 Monday
##  8   1400     898      1 Monday
##  9   1800     385      1 Monday
## 10   1400     416     25 Monday
```

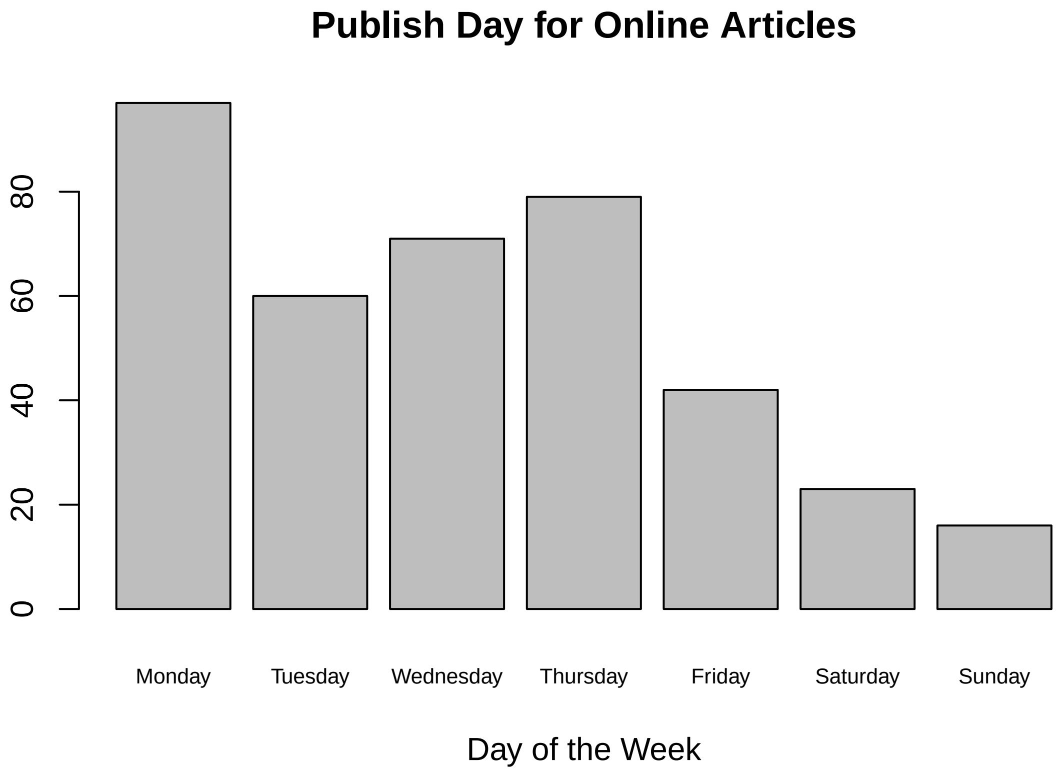
Univariate EDA

We examine the 4 variables individually by their appropriate histograms (shares, content, and images) or barplots (day published):


Number of Shares For Online Articles
Frequency
0
20
40
60
80
500
1000
1500
2000
2500
3000
Number of Shares


Number of Words in Online Articles
Frequency
0
50
100
150
200
0
500
1000
1500
2000
2500
3000
3500
Number of Words


Number of Images in Online Articles
Frequency
0
50
100
200
300
0
10
20
30
40
50
60
Number of Images



In addition, we provide the numerical summaries for the individual variables:

Shares:

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

Content:

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

Images:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     1.0     1.0     4.4     3.0    61.0
```

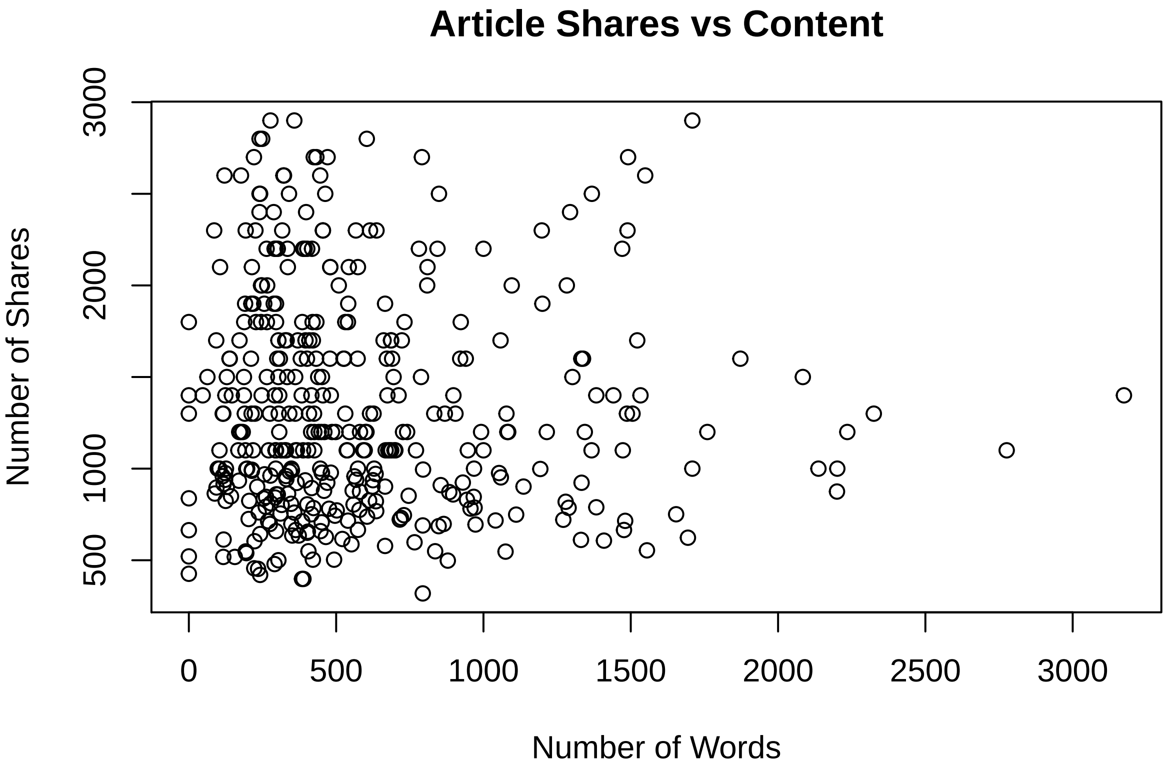
Day Published:

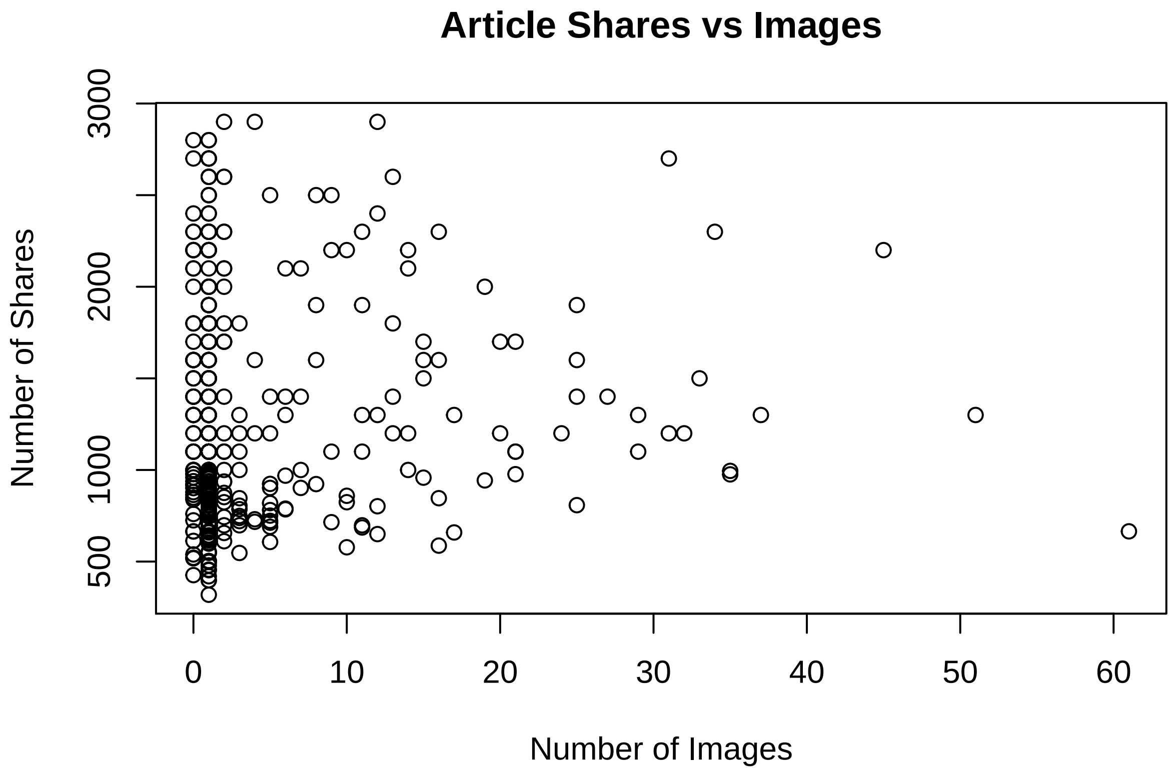
```
##    Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
##        97        60        71        79        42        23        16
```

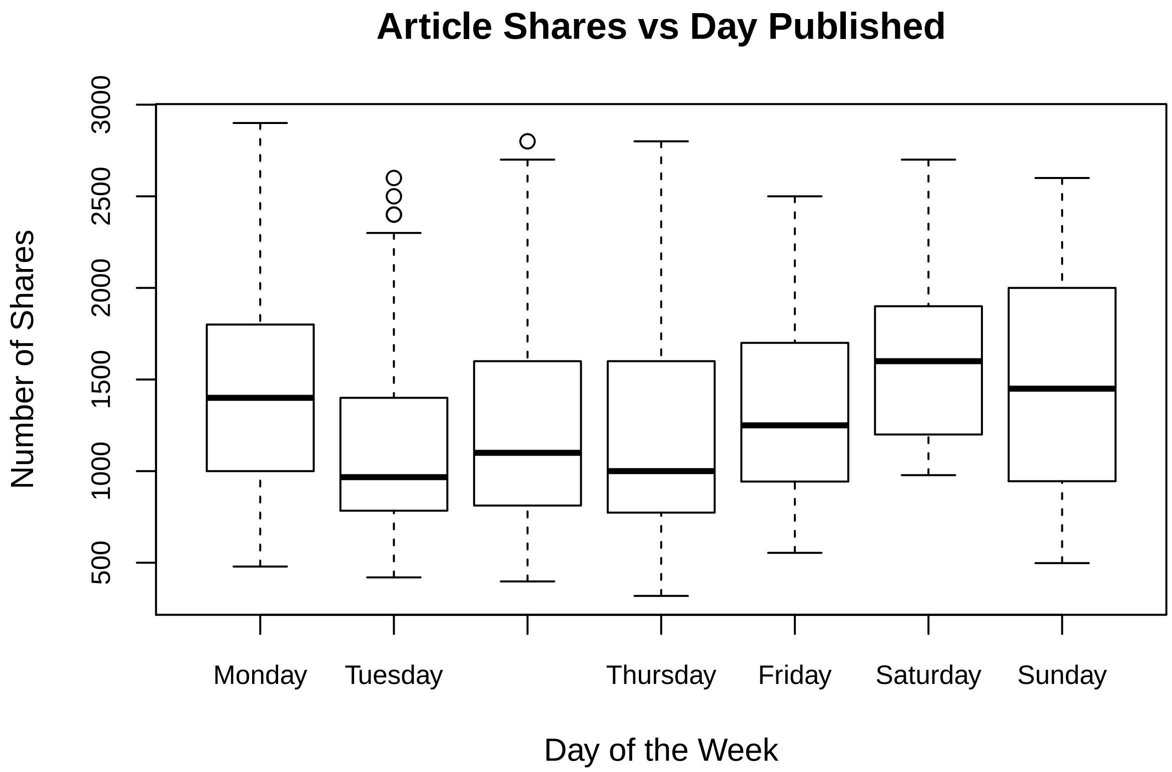
For the variable **shares**, we observe a unimodal and approximately right-skewed distribution. It has a mean of about 1325 and a median of 1200. For the variable **content**, we once again observe a right-skewed and unimodal distribution. There may be potential outliers at the tail of the distribution. We also see a relatively wide range (0 to 3174) for the number of words in articles. For the variable **images**, we see that most online articles have 0-5 images, with very few articles that have more than the mean number of images, which is 4.4. For the variable **day published**, we notice that Monday has the most online articles published (97 articles), whereas Sunday has the least online articles published (16 articles).

Bivariate EDA

After inspecting the individual variables in the dataset, we now observe the relationship between the predictors and the response variable (shares), as below:




Article Shares vs Images
Number of Shares
500
1000
2000
3000
0
10
20
30
40
50
60
Number of Images


Article Shares vs Day Published
Number of Shares
500
1000
1500
2000
2500
3000
Monday
Tuesday
Thursday
Friday
Saturday
Sunday
Day of the Week

```
## [1] -0.02593102
```

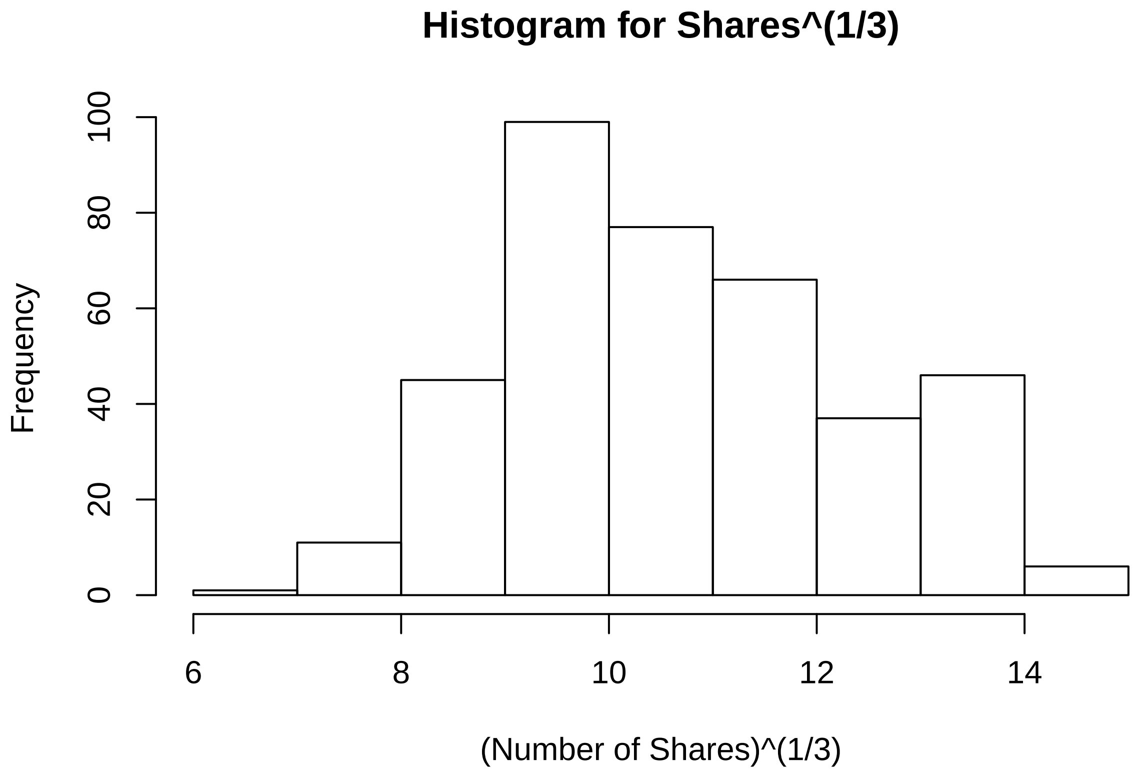
```
## [1] 0.04562069
```

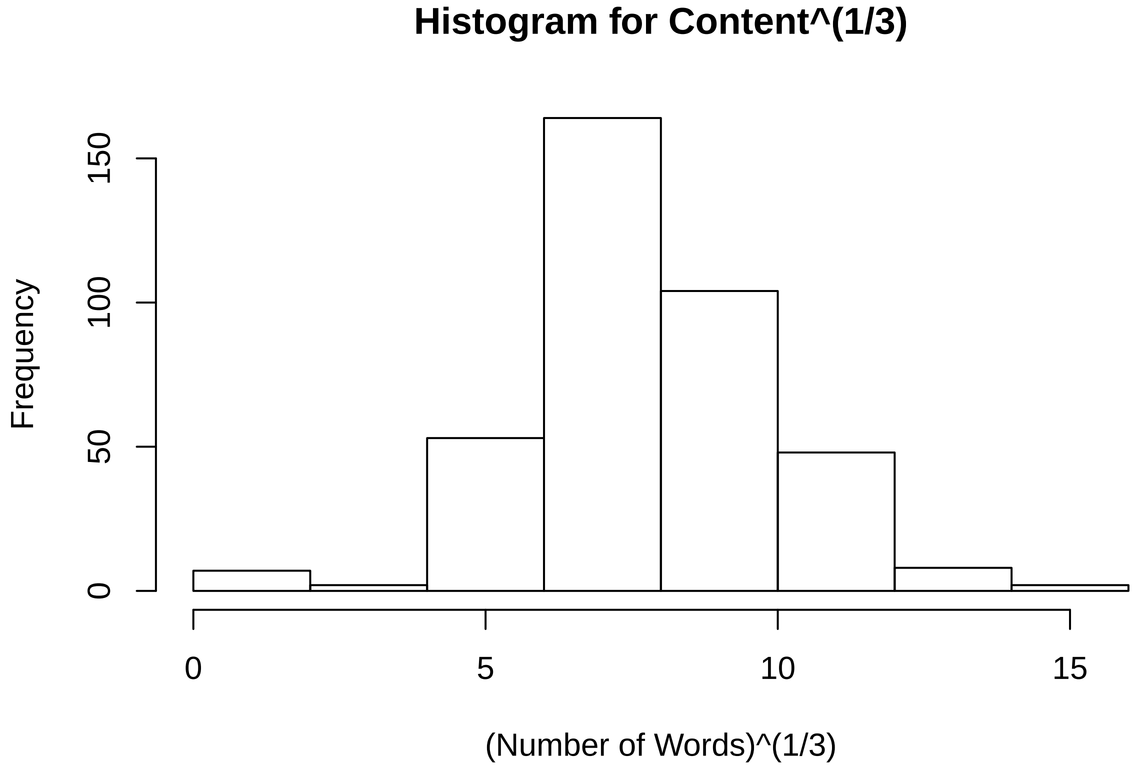
Based on the graphs above, the positive/negative relationships isn't very clear between the variables **shares** and **content** (around -0.03 for their correlation coefficient). For the variables **shares** and **content**, we see a very weak positive relationship (correlation coefficent around 0.05). In both scatterplots, it seems like we can somehow fit a straight line, indicating that they might be linear. After examining the boxplot, we notice some slight variations in article **shares** among the different **day published**. However, the differences in the number of shares among which day of the week is also weak, since there is much overlap between all the boxes.

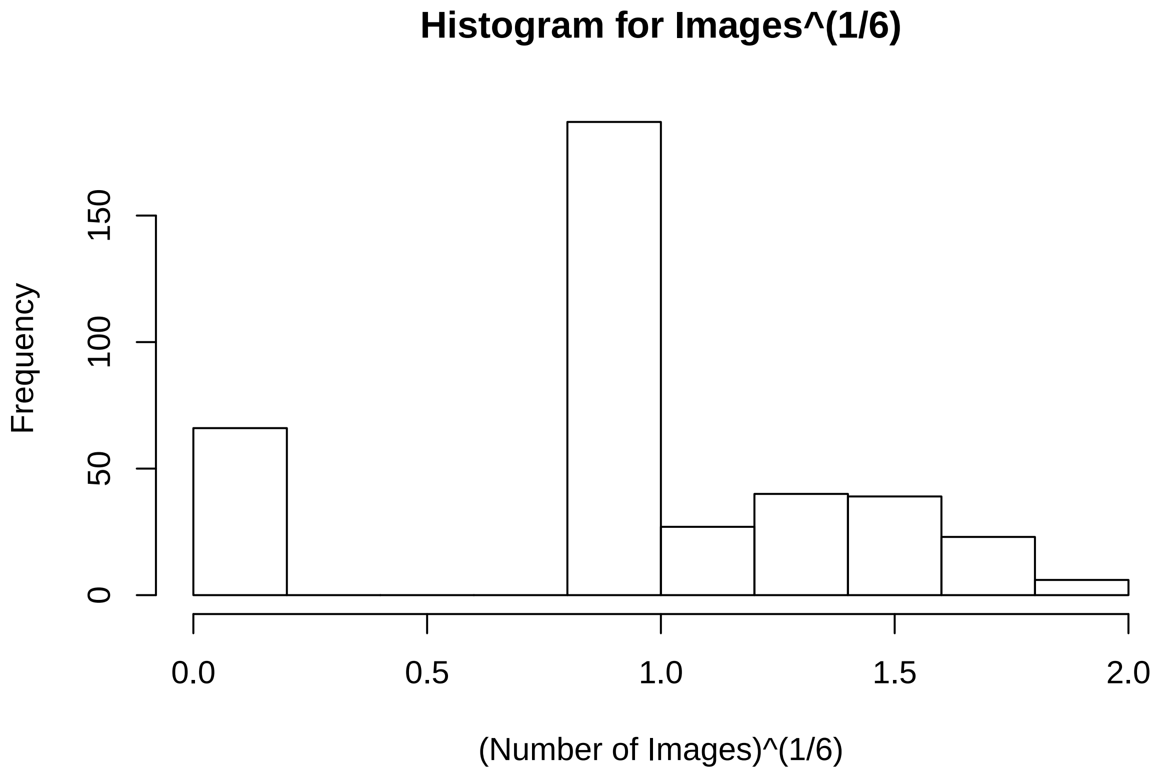
Modeling:

Transformation

According to the univariate EDA, we see that the histograms for all the variables are severely right skewed and therefore not symmetric. As a result, we need to transform the data to improve the multiple linear regression model diagnostics and to better predict the number of shares. In this case, we apply a cubic root transformation on the response variable **shares** and the predictor **content**, as well as a 6th root transformation on the other quantitative predictor **images** since they seem to be the best transformations to normalize the histograms without fixing the data too much, as shown below:


Histogram for Shares^(1/3)
Frequency
0
20
40
60
80
100
6
8
10
12
14
(Number of Shares)^(1/3)

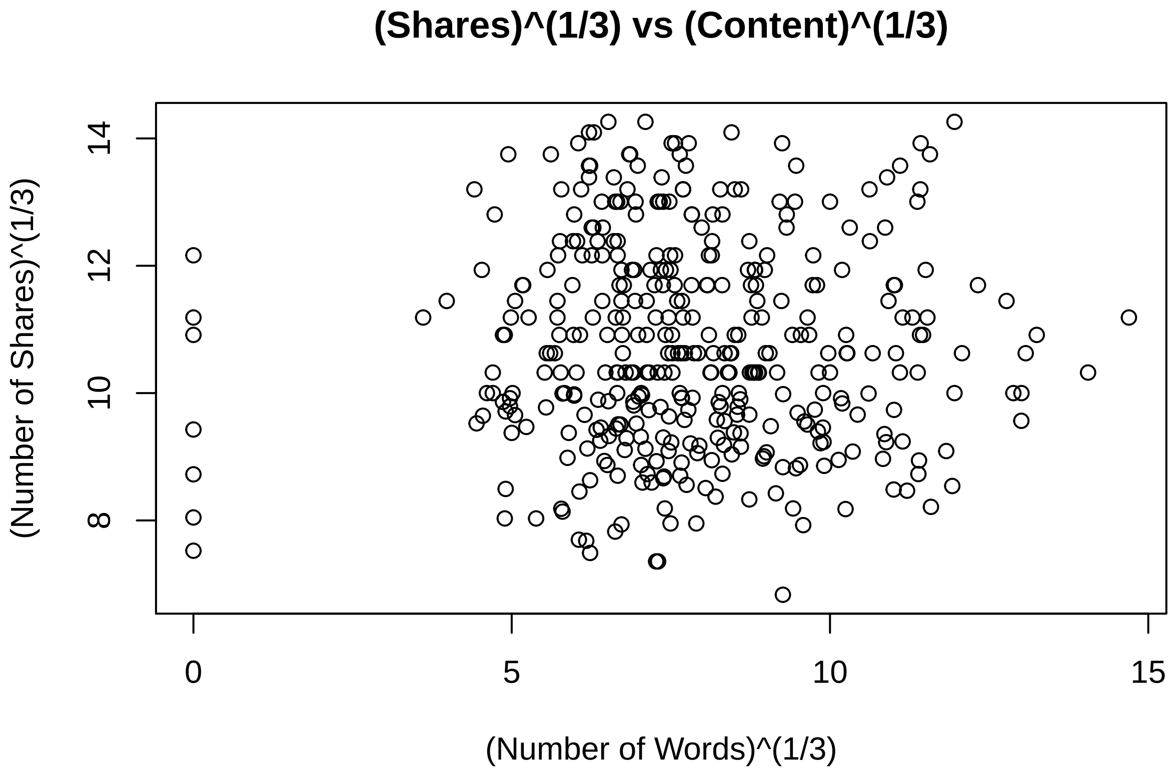

Histogram for Content^(1/3)
Frequency
0
50
100
150
0
5
10
15
(Number of Words)^(1/3)

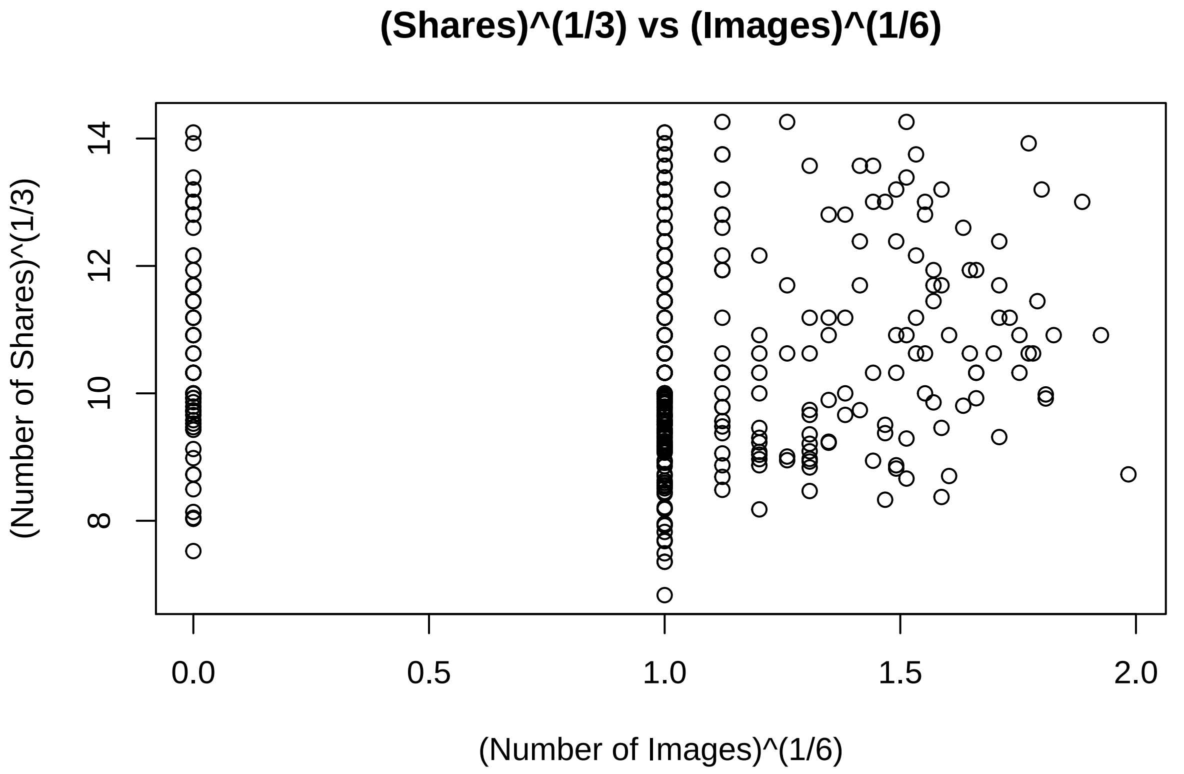


Linear Assumption

After numerous trials of finding the best transformation for the severely right-skewed histograms, we obtain some new scatterplots on the transformed variables. The linearity assumptions seem to have slightly improved (more spreaded out), but the relationships between the Y variable (shares) and the X variables (content, images, day published) are still very weak (as seen in the scatterplots below and the bivariate EDA). Despite the weakness, the linear assumption is still justified since we don't see any curves in the graphs and fitting a straight line seems probable.

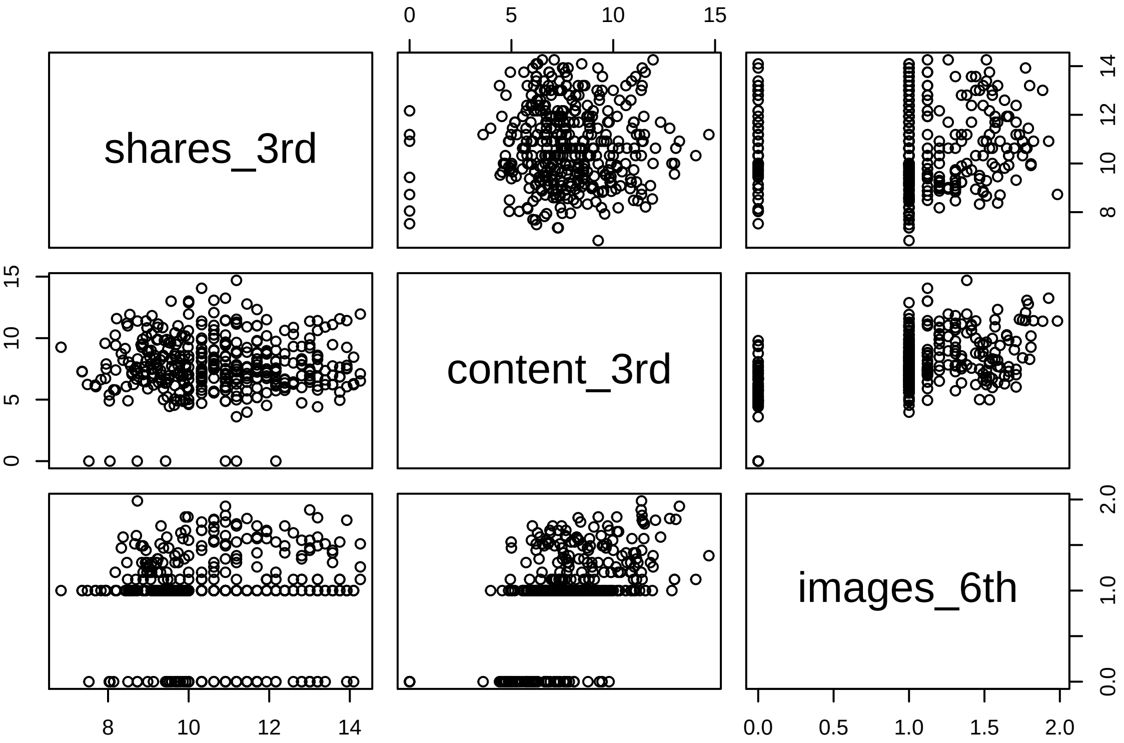
Note: The stacked data appears on the scatterplot for Shares vs (Images)^(1/6) no matter which appropriate transformation for remediating right-skewed data is applied. This transformation is the best solution we can find after deliberate consideration.


(Shares)^(1/3) vs (Content)^(1/3)
(Number of Shares)^(1/3)
8
10
12
14
0
5
10
15
(Number of Words)^(1/3)



Multicollinearity

To see if the variables can be used in the multivariate linear model, we now check for multicollinearity for the quantitative variables. We use the pairs plot to observe if there's a strong relationship between the predictors. We do not include **Day Predicted** in the pairs plot since it is a categorical variable.

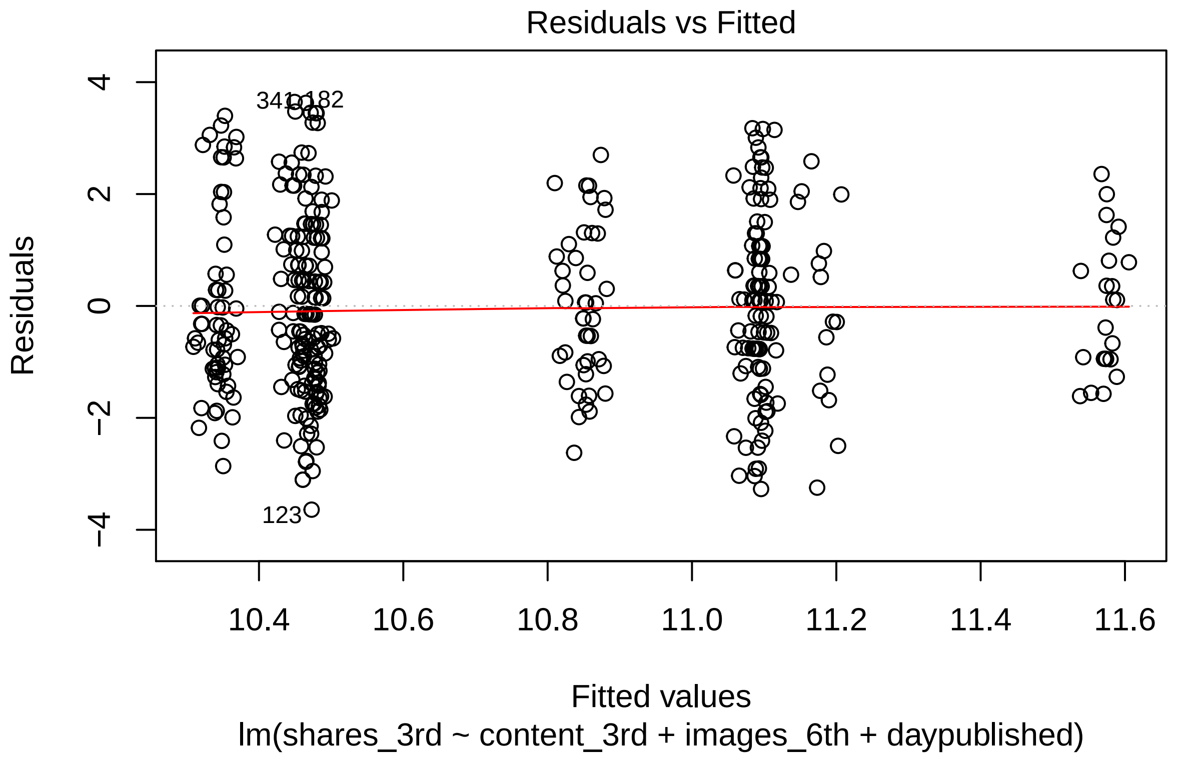


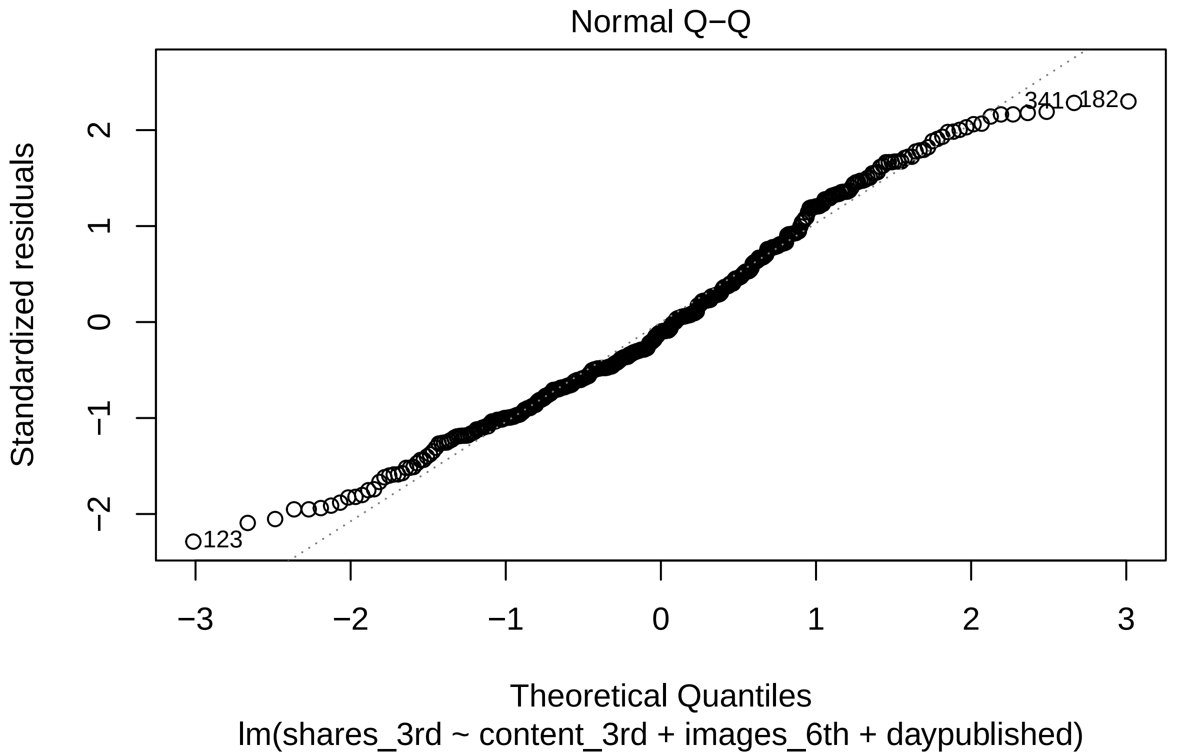
Based on the pairs plot above, we don't see any strong relationships between any of the predictors. We can check this again with a more reliable diagnostic, which is by analyzing the variation inflation factors (vif) for the same variables:

```
##                   GVIF Df GVIF^(1/(2*Df))
## content_3rd  1.375954  1        1.173011
## images_6th   1.372717  1        1.171630
## daypublished 1.020058  6        1.001656
```

By the nature of this model, we obtain GVIF's in this case, but they are interpreted in the same manner as vifs. Since the GVIF's for all three predictors are less than 2.5, we can conclude that this model is not in danger of multicollinearity (in other words, this model meets the multicollinearity model assumption). As a result, we can make the residual and QQ plots based on these variables.

Residual and QQ Plot





The residual plot above validates the following error assumptions about the model:

- independence: the residuals are patternlessly and randommly scattered above and below the zero line
- zero mean: the residuals are centered around the zero line
- constant standard deviation: as we inspect across the residual plot, we see that most of the residuals are constantly spreaded above and below the zero line

The QQ plot above tells us whether the residuals are normally distributed. While there are few deviations at the two ends, most of the points are very close to the line on the plot. Thus, the QQ plot still justifies the assumption that the residuals are normal.

Model Summary

We supplement the multiple regression analysis summary for our model below:

```
##
## Call:
## lm(formula = shares_3rd ~ content_3rd + images_6th + daypublished,
##     data = social)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6401 -1.1191 -0.1645  1.0977  3.6452
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.847448   0.375309  28.903   <2e-16 ***
## content_3rd           -0.003979   0.043200  -0.092    0.927
## images_6th             0.035365   0.192116   0.184    0.854
## daypublishedMonday     0.239326   0.296150   0.808    0.420
## daypublishedSaturday   0.721540   0.416536   1.732    0.084 .
## daypublishedSunday     0.328761   0.472208   0.696    0.487
## daypublishedThursday  -0.373063   0.306294  -1.218    0.224
## daypublishedTuesday   -0.507595   0.322579  -1.574    0.116
## daypublishedWednesday -0.393072   0.312418  -1.258    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.603 on 379 degrees of freedom
## Multiple R-squared:  0.05085,    Adjusted R-squared:  0.03082
## F-statistic: 2.538 on 8 and 379 DF,  p-value: 0.01058
```

This model is justified for predicting **shares** since both the linear and error assumptions for multiple linear regression models are validated (as seen previously). When we tried different models that included interaction terms between **day published** and the other predictors, some model assumptions were violated. We also observe that the p-values for interaction terms are all greater than 0.05, which implies no significant interaction between the predictors. Some models with interaction terms have R^2 values slightly lower and higher than what we have in this model, but the model with higher R^2 violated the multicollinearity assumption (some GVIF's were greater than 2.5). To balance between meeting the model diagnostics and obtaining as high as the R^2 can be (about 5.1 in this case), this is the best model we can obtain with the data we're given and the transformations we did.

We notice that the model is significant since the p-value for this regression F-test is 0.01058, which is less than 0.05; however, the individual predictors are not significant in this model. We tried various subsets of the model (regressing Y on one predictor, two predictors, three predictors at a time, and the possible combinations of adding interaction terms), which all result in the individual p-values greater than 0.05. Hence, there is sufficient evidence that our model is signicant, but not the individual variables.

We also observe a negative coefficient for content^(1/3) and a positive coefficient for images^(1/6). These values validates the relationships we have seen in the bivariate EDA. Moreover, we have 6 dummy variables for the variable **day predicted**. Tuesday, Wednesday, and Thursday have negative coefficents, which confirms with the boxplots results above (lowest medians), while Monday, Saturday, and Sunday have positive coefficients (highest medians based on the boxplots).

Prediction

Despite reaching the conclusion that the predictors **content**, **images**, and **day published** are insignificant in terms of their p-values, we decide not to leave them out because we might be accidently removing other important information that are useful for us. Also, we now understand that these factors do not play a significant role in predicting the number of online article shares, but it doesn't mean that our model is bad in any way. Moreover, the predictors may be deemed important theoretically.

After obtaining a reasonable model for predicting **shares**, we now use it to answer the client's queston, which is to predict the number of shares for an online article that has 627 words, 3 images, and was published on Saturday:

Note: The coefficients for other dummy variables are not included since they have x values equal 0; the dummy variable **daypublishedSaturday** has 1 for Saturday.

```
(10.847448 -0.003979*((627)^(1/3)) + 0.035365*((3)^(1/6)) +0.72154*1)^3
```

```
## [1] 1551.792
```

We conclude that the predicted number of shares is around 1552 when an online article is published on Saturday with 627 words and 3 images. It is slightly higher than the overall mean (1325), and within one standard deviation away. However, since the predictors are insignificant, this may not be the most accurate equation to predict the number of online article shares.

Discussion

After analyzing this social media dataset from Mashable, we observed that the number of words, images, and the day (in a week) when an article is published are not linearly related to the number of article shares. However, the F-test tells us that there is significance when considering the variables jointly. Regarding the model assumptions, we transformed the response variable and the two qualitative predictors to improve the linear assumption diagnostic. We did not encounter any multicollinearity issues in this sample. The error assumptions were confirmed by the residual diagnostic plot and the QQ plot. We noticed some outliers at the tails of the QQ plot, but they did not affect much about the overall result.

Other potential predictors to consider for predicting the number of article shares might be the time when an article is first posted, which social media is used, or the different categories of news. It is important that we continue investigating the possble indicators that would result in higher article shares as more people spend more time consuming informaton via the internet. This would benefit companies that make money based on the number of views they receive for the articles they publish, or other indiviuals who desire to spread information in a faster manner.