Predicting Shares of Articles on Social Networks

Xuduo Victor Wen

Contents

Introduction

The power of social media often lies in its ability to spread information (correct or not) through networks of readers. Readers share articles online, and questions therefore arise about what features might predict whether online content gets shared. In the present paper, we have the available data of a sample of online articles published on the site Mashable over a period of two years. We will focus on the number of shares the article has in social networks and determine whether there are specific factors that may contribute to increasing/decreasing the number of shares for any articles.

Exploratory Data Analysis

Data

In this social media data, we analyze a random sample of 388 articles and 4 variables. Due to our interest in factors that contribute to the number of shares the article has in social networks, we examine the relationship between the number of shares, our response variable, and three explanatory variables: content, image, and the day of publication. We summarize the variables as follows:

shares: the number of shares the article has in social networks (measured on a quantitative index, with larger values representing more shares).

content: the number of words in the article (measured on a quantitative index, with larger values representing more words).

images: the number of images in the article (measured on a quantitative index, with larger values representing more images).

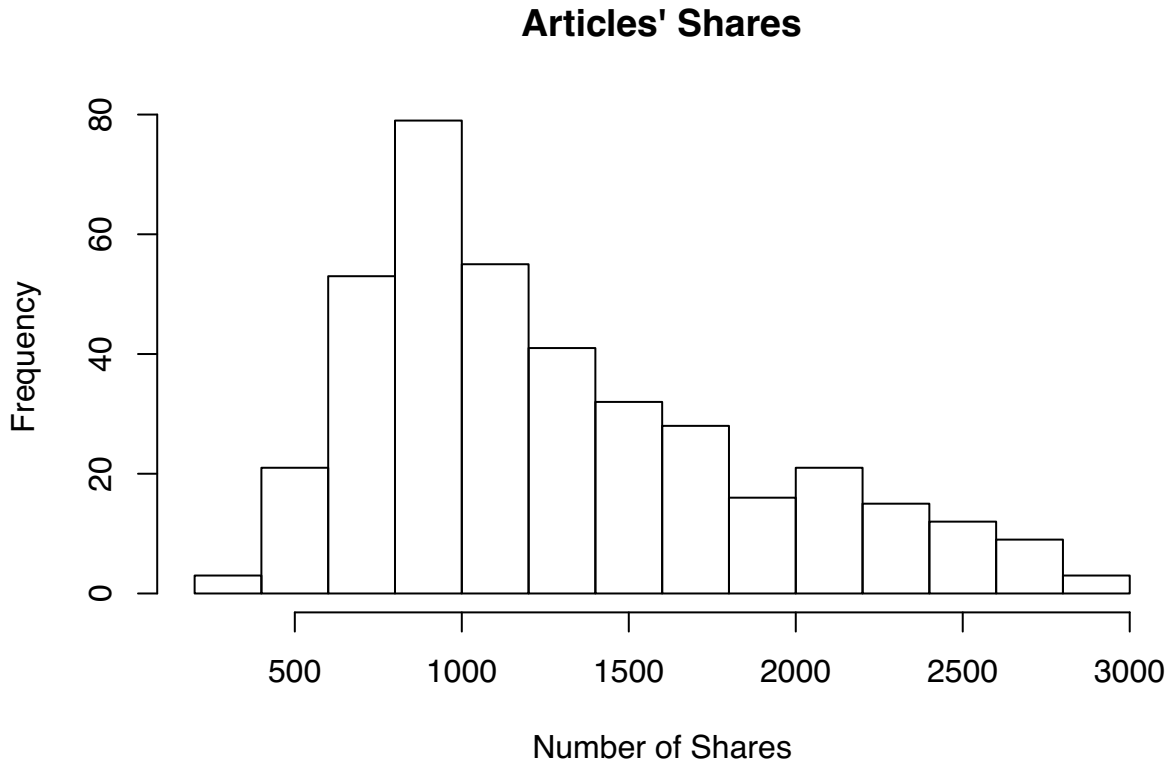
daypublished: the day of the week the article was originally published (measured on a categorical index, with 7 different values indicating days oof the week. For example, Monday, Tuesday. . . .).

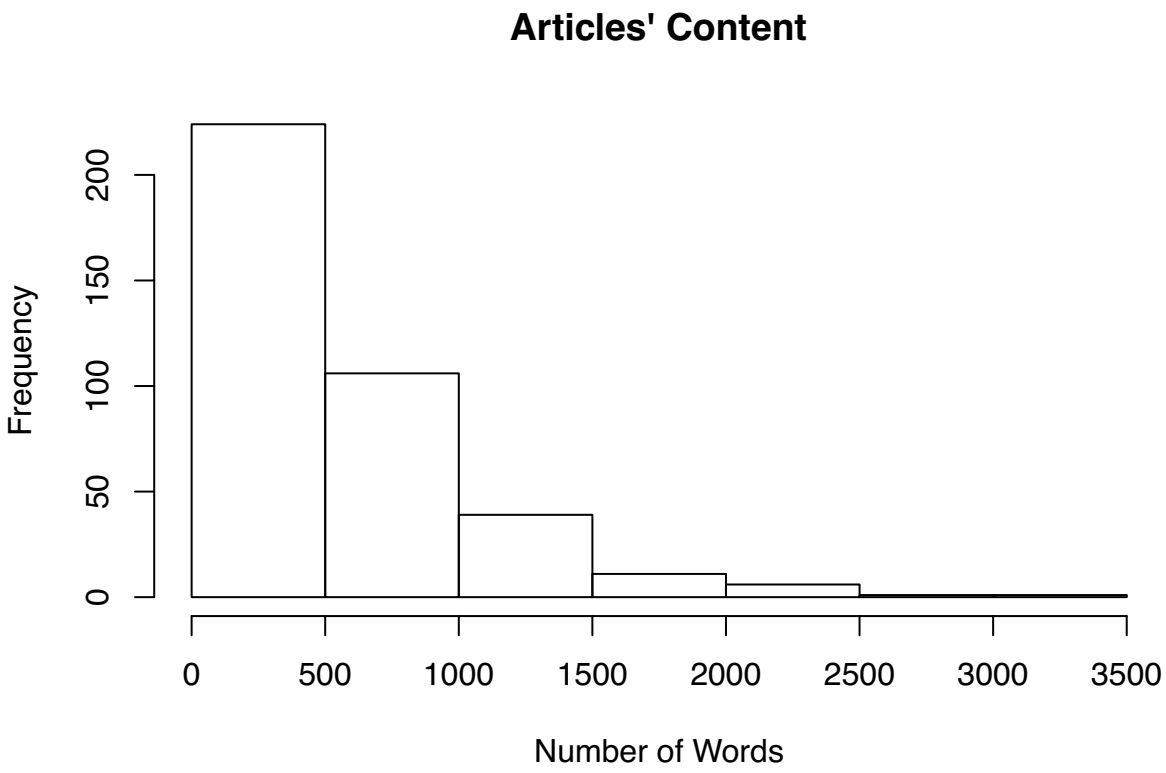
The first few lines of data appear as follows:

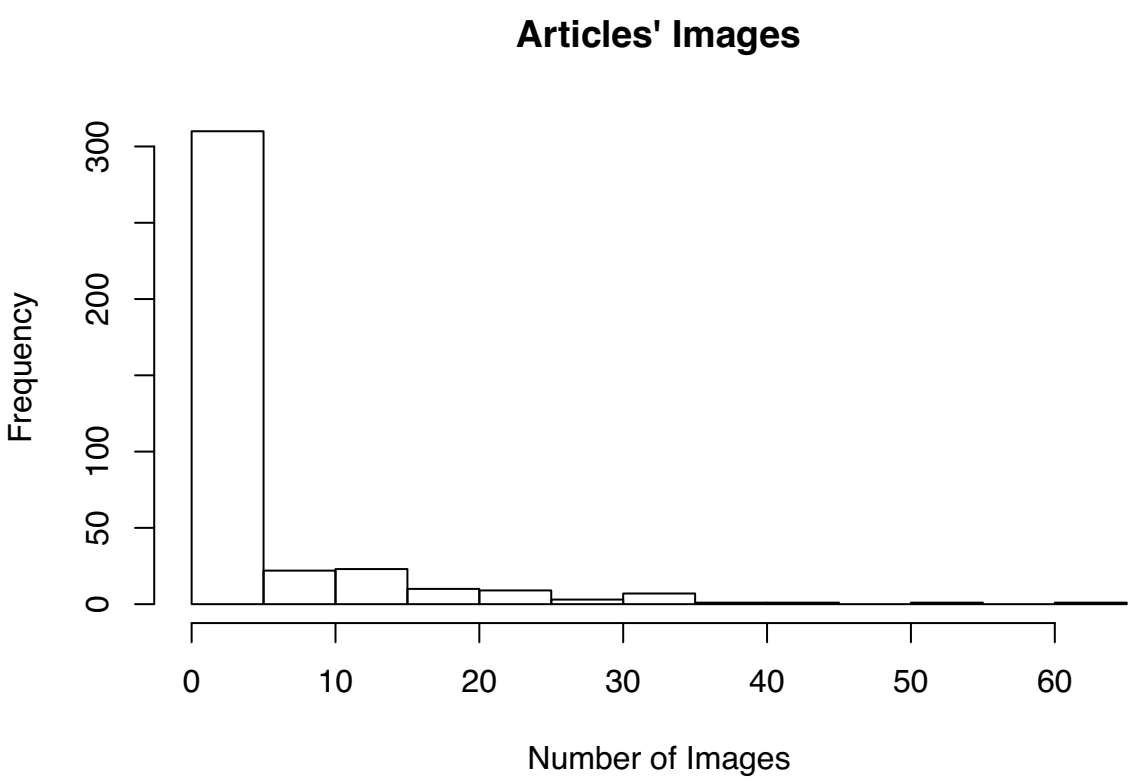
```
## # A tibble: 6 x 4
##   shares content images daypublished
##    <dbl>   <dbl>  <dbl> <chr>
## 1   1100     367      1 Monday
## 2   1400     712      1 Monday
## 3    479     291      1 Monday
## 4   2500     463      5 Monday
## 5   1200     498     13 Monday
## 6   1200    1084      1 Monday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

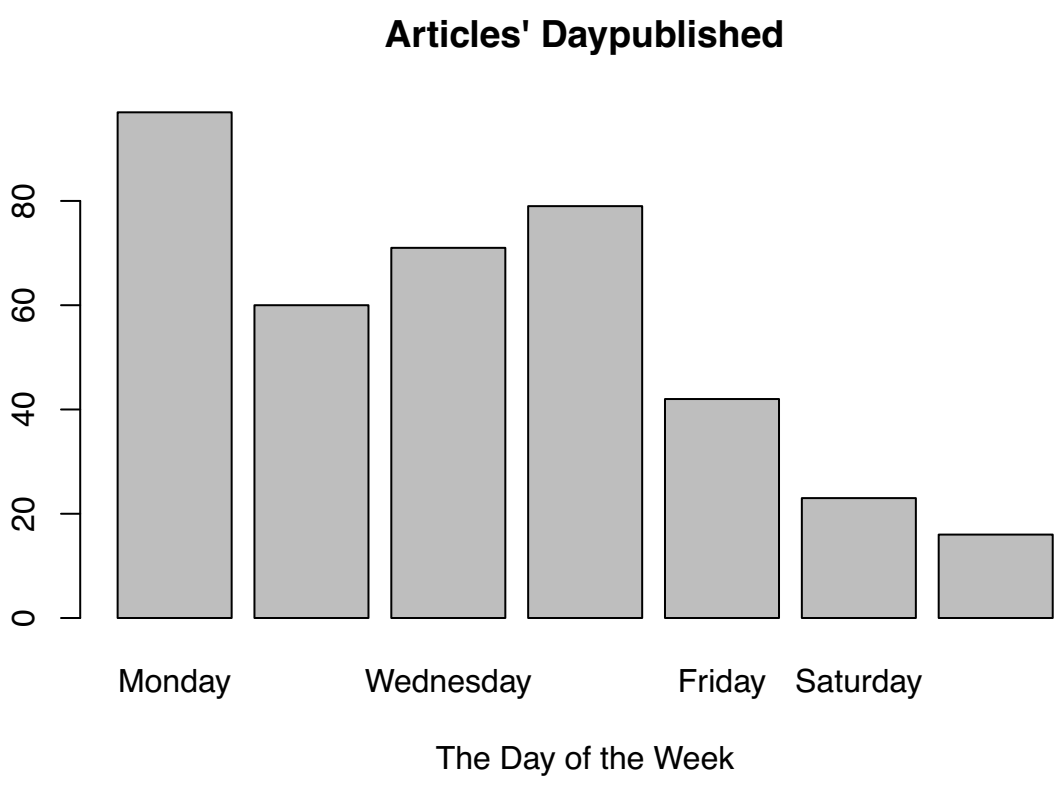
Univariate Exploration

As a first step in the analysis, we explore each variable individually. We use histograms to explore the distribution of continuous variables and a barplot to explore our categorical variable.




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


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



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

For Shares

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

```
## [1] 598.5999
```

For Content

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

```
## [1] 470.5251
```

For Images

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

```
## [1] 8.211868
```

For Daypublished

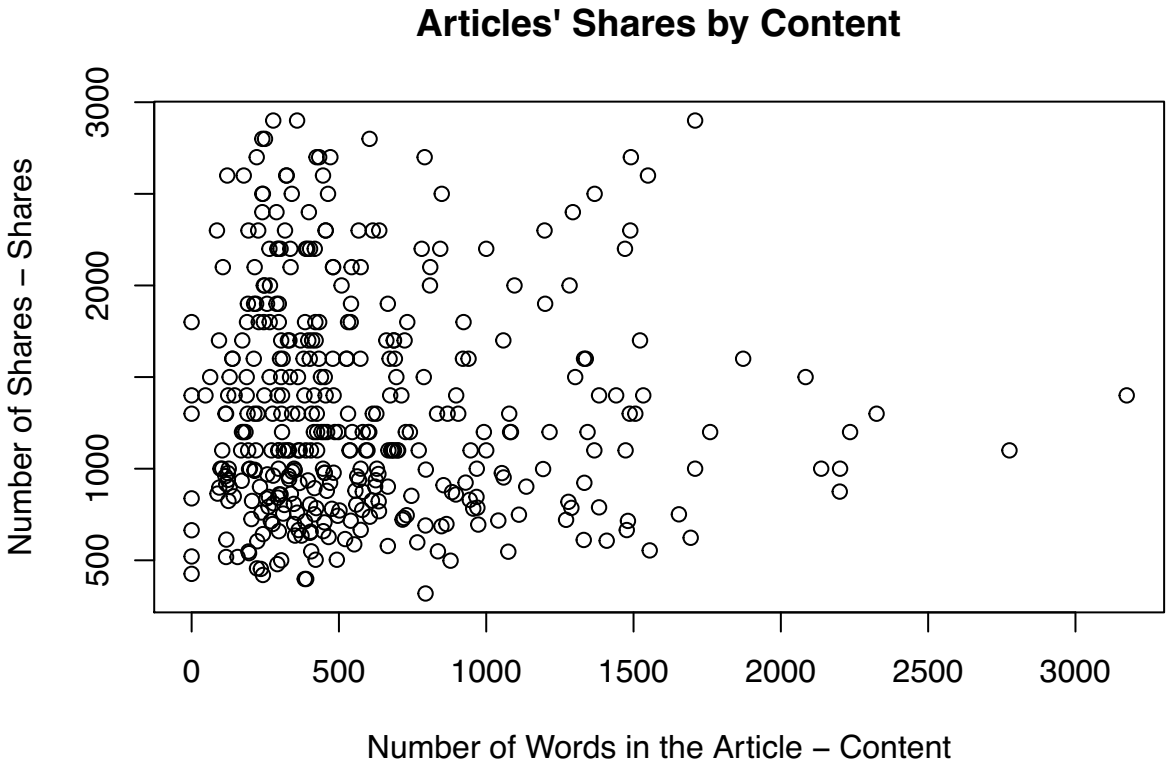
```
##    Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
##        97        60        71        79        42        23        16
```

After looking at both the graphs and the summary statistics of our variables, we make the following observations: The distribution of article's **shares** is approximately unimodal and right skewed. This would indicate that the median is a bit smaller than the mean, as our sumerical summary confirms. The distribution is centered around 1325 and the standard deviation is around 600. The distribution of articles' **content** is

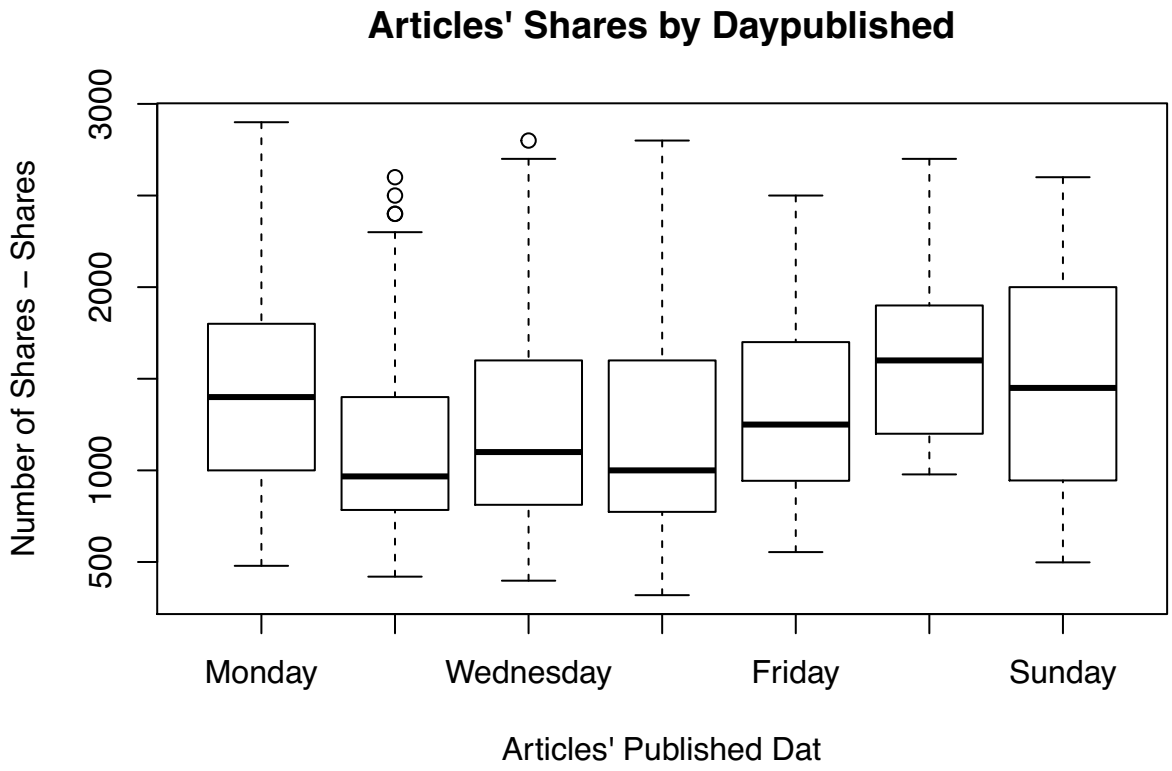
very much right skewed and unimodal. From the histogram, we can see the peak in the variable on the far left of the graph. The distribution is centered around 433 with the standard deviation around 470. The number of **images** in the articles range from 0 to 61 with an average of about 4.433. Most articles have roughly less than 3 images (majority of 75%), with a few articles that have a lot of images. The distrubution is heavily right skewed. The outliers are significant in the sense that the standard deviation of the distribution is around 8. Looking at the articles' **publication day**, we find that there are Mondays have the most published articles in our sample, and the weekends have the least published articles in our sample. The distribution of daypublished is roughly bimodal with two peaks on Mondays and Thursdays. Overall, our four histograms are all somewhat right skewed.

Bivariate Exploration

Now that we understand the distribution of the individual variables in this data, we can graphically how each predictor is associated with the response **shares**, as follows:




Articles' Shares by Images
Number of Shares – Shares
500
1000
2000
3000
0
10
20
30
40
50
60
Number of Images in the Article – Images



Through analysis of our graphs, we find that articles' **shares** is negatively, and have weak linear association with **content**. As the number of words increases, the number of shares satisfaction usually decreases except a few distinguished outliers at around content = 1500. We have a similar observation on association of **shares** with **images**. There are an abundace of plotted points for articles that have less than 10 images. The association is slightly positive and weak linear association as well. The two scatterplots above share a lot of similarities. In general, despite the low number of total shares on weekends, the median numbers of shares are larger on the weekends than all of the weekdays. Monday has the highest median shares of all weekdays. There are 3 outliers above the maximum of shares on Tuesday and 1 on Wednesday. The ranges of shares on the weekends are also lower than the ranges of shares on the weekdays. If we were to examine the boxplots starting from Saturday and ending on Friday, we can somewhat state that there's a decreasing trend of **shares** on **daypublished** strictly for the boxplots.

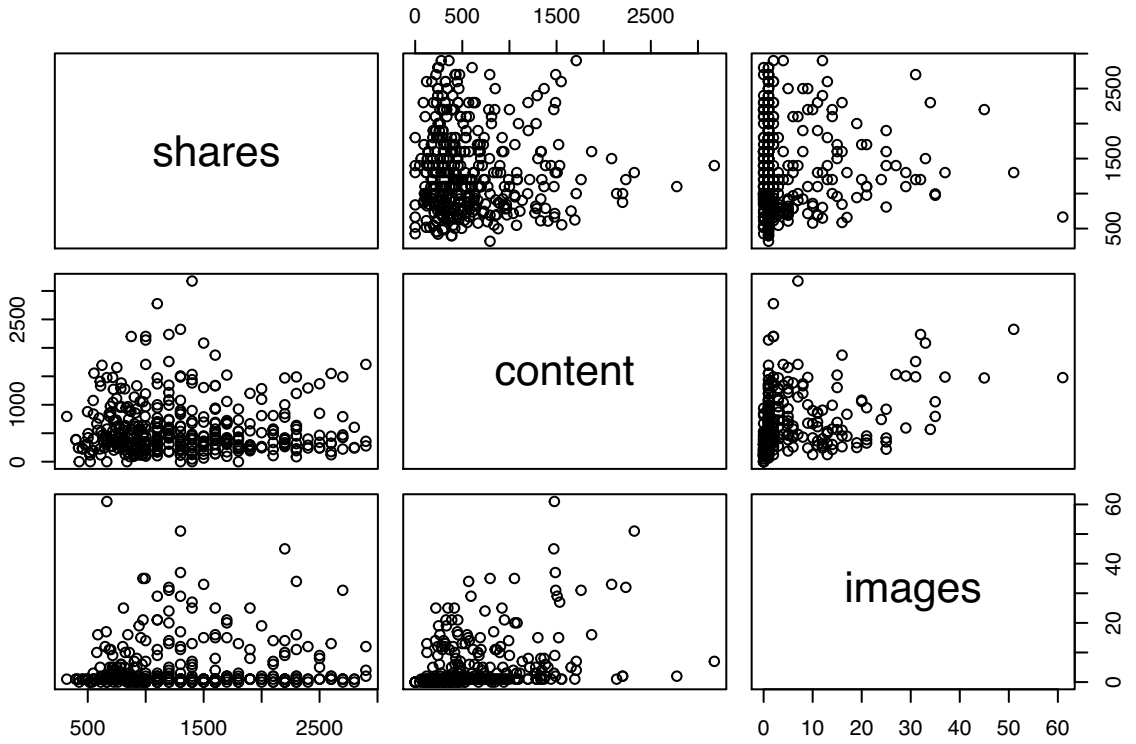
Modeling

After exploring and visualizing the relationships among our variables, we now turn to building a linear regression model to predict articles' shares. We start by looking at the histogram of our response variable. It looks a bit right skewed, indicating that a transformation might be needed. For now, we leave the response variable in its current form and will comment on potential transformations back when looking at model diagnistics.

We saw in our bivariate exploratory data analysis that all variables have some relationship (either weak or strong) with articles' shares. Therefore, all variables may be useful in this model, but we need to first check for multicollinearity. A first indication of possible multicollinearity is relatively strong correlations between pairs of explanatory variables, so we can check the pairs plot.

[We remark that for the purpose of the pairs plot we first subsetted out the categorical variable **Daypublished**.

However, if we choose to examine the categorical variable with other quantitative variable as predictors, then we would need to use a multiple linear regression model with potential interaction terms If such interaction(s) between categorical and quantitative predictor were significant, from looking at the p value(s) of the interaction terms, then we would keep all predictors that are in the interaction.]




```
##         shares content images
## shares    1.00   -0.03   0.05
## content  -0.03    1.00   0.37
## images    0.05    0.37   1.00
```

The relationship between **shares & content, and shares & images** are extremely weak. From the correlation coefficient (r), we observe that the r values are -0.03 and 0.05. On the other hand, there is a relatively observable positive association between content and images, so we might be concerned about multicollinearity and may not want to include both variables in our model. We formally check the variation inflation factors (vif) for these variables in a full multilinear regression model predicting **shares** from each of the three predictors:

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

[We parenthetically remark that, depending on the type of variables and on the nature of the model built, we might have obtained "GVIF" from the vif() function, and if so, that GVIF's are interepreted the same way as vif's.]

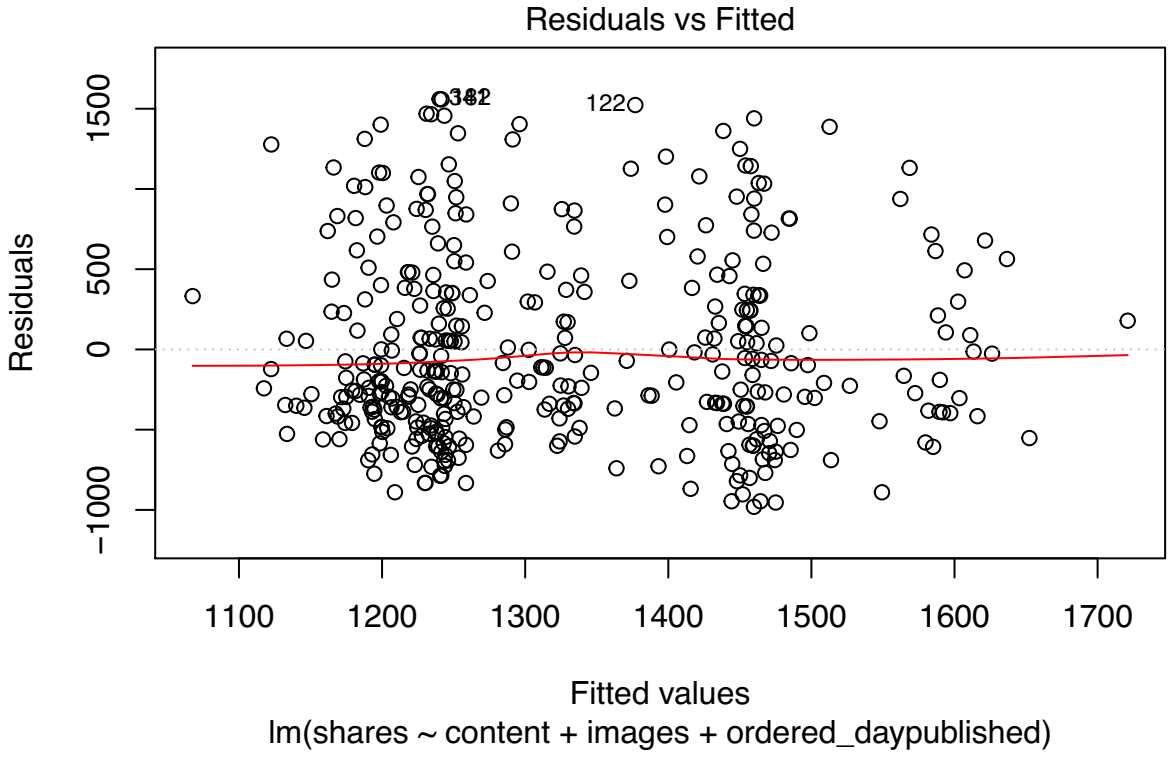
We see from above that all vif values are less than 2.5 so our multicollinearity model assumption has been met. Having obtained a model without dangerous multicollinearity, we demonstrate the residual diagnostic

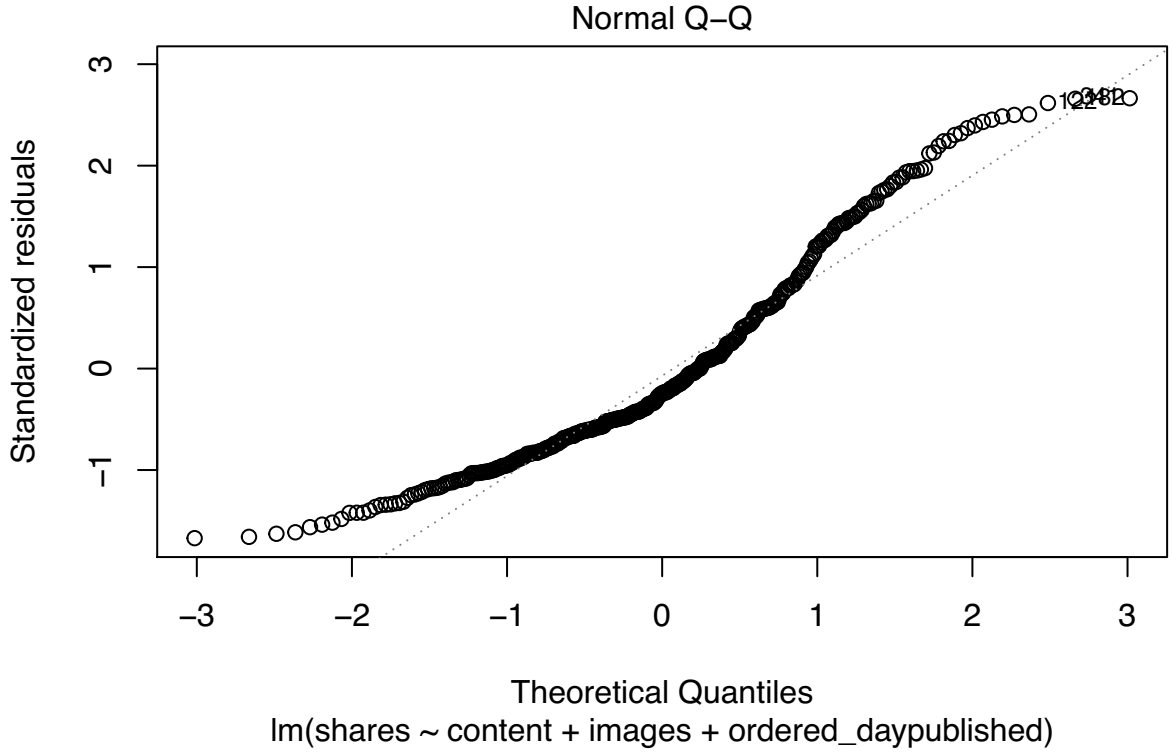
plots (below) from our new model with **content,images, and daypublished**.

On the residual plot, we oberve that aside from a few outliers with notably larger residual (from row 122 and others as indicated), the constant spread, independence, and mean zero assumptions are otherwise as reasonably justified as we could obtain (because the residuals have roughly equal spread above and below the 0 line and also do not present any obvious pattern).

On the qqplot, we note some deviation on the ends as well as the same outliers (from row 122 and others) on the upper end; on balance these do not seem too severe to invalidate the normality condition since the rest of the points fall reasonably near the line on the qqplot; and this was about the best normality diagnostic we could obtain of the various models we tried to build.

We note that the residual assumptions are reasonably satisfied, and hence we encountered no need to try transformations.(However, if we apply more rigorous standard on the normal qq plot and deem it somewhat invalid, we could potentially try a logarithmic transformation on the linear model that could potentially result in a better fitting model. Such tranformation could be considered in future scenarios.)





With the prior remarks an analysis in mind, the regression analysis summary from fitting our final chosen model is as shown; discussion follows:

```
##
## Call:
## lm(formula = shares ~ content + images + ordered_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)                   1475.06814   70.20663  21.010  < 2e-16 ***
## content                         -0.06846    0.06905  -0.992  0.32204
## images                           4.71893    3.98135   1.185  0.23666
## ordered_daypublishedTuesday   -268.62326   97.27195  -2.762  0.00603 **
## ordered_daypublishedWednesday -223.18408   92.31995  -2.418  0.01610 *
## ordered_daypublishedThursday  -216.43919   89.48316  -2.419  0.01604 *
## ordered_daypublishedFriday    -127.26428  109.32710  -1.164  0.24513
## ordered_daypublishedSaturday   142.99237  137.10810   1.043  0.29765
## ordered_daypublishedSunday      24.63381  159.90302   0.154  0.87765
## ---
## 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
```

We consider this a reasonable model to predict shares, since all quantiative predictors showed a linear relationship with articles' **shares** (as seen through our EDA); hence, the model do justify the linearity condition of the multiple linear regression model. We see that the interaction model is significant since there is at least one significant interaction term. We see that the model is significant as indicated by the regression F-test p-value of 0.01562, which is less than our presumed alpha level of significance (5%).

We remark that we see negative coefficient value associated with content and positive coefficient value associated with images, which confirms our EDA results. We see positive coefficient values of **ordered_daypublished Saturday and Sunday** associated with the categorical dummy variable **ordered_daypublishedMonday**, which reflects our results from the side-by-side boxplot: weekends, on average, have higher number of shares. We see negative coefficient values from **ordered_daypublished Tuesday to Friday** associated with the categorical dummy variable **ordered_daypublishedMonday**, which reflects our results from the side-by-side boxplot: weekdays, on average, have lower number of shares.

In this linear regression model there is an acceptable minimum of multicollinearity (indicated by the VIFs $<$ 2.5), the signs of the coefficients are consistent with the individual simple regressions and our EDA, and this model has as high an R^2 as could be reasonably found while balancing residual diagnostics and relative simplicity. We are confident that articles that have lower number of words, greater number of images, and published on weekends rather than weekdays are associated with larger number of shares.

Prediction

Now that we have a model that reasoably satisfies all assumptions, we are interested in predicting the number of shares for an article that has 627 words, three images, and was published on Saturday.

The predicted satisfaction is computed as follows:

1475.06814 - (0.06846 x 627) + (4.71893 x 3) + 142.99237 =

```
## [1] 1589.293
```

[We remark the software-created dummy variable "Ordered_daypublished Saturday" is understood to be 1 for Saturday; it would be 0 otherwise]

The predicted shares of an article, that has 627 words, three images, and was published on Saturday, is roughly 1589. [We remark that this value is considered a relatively higher number of shares, as it is within the top 25-50% of all shares.]

Discussion

In this analysis, we learned that articles' number of shares on social networks is related to the number of words (content) in the article, the number of images in the article, and the day published. In the model, there are still insignificant predictors so it could be perfected, if we were predicting shares of an articles with different predictors (since we were given images, content, and day published, we will stick with this model).

However, if we were using the same scenario in the future without the necessity to utilize all predictors (from the potential customer), we can use a reduced model of only significant predictors (only dayspublished in this case) and compare the R^2 value of such model to our current model. The greater the R^2 value is, the more likelihood that greater amount of variation could be covered by the potential model.

We noted some issue with some outliers (evident in the residual diagnostic and qq plots); it could be useful to investigate these individual values further if possible.

The relative difference in shares among different days of the week is notable, and could be an area for further investigation. We further note that social media platform information is missing from our available data; we might like to know, for instance, if different social media platforms are included and what role they may play in increasing or decreasing shares.

Overall, data analytics will continue to play an ever more vital role on social media. Platforms should continuously evaluate the different characteristics of any largely shared articles to increase user engagement. Analyses like these are beneficial to the regulators, content creators as well as users.