What drives the sharing of Mashable Articles?

Dylan Chou
dvchou

11-28-18

Introduction

With the sweeping wave of new technologies and increasingly interconnected online networks, sharing has become easier than ever before. However, with the large volume of users on social media, it's important to determine what drives people to share the things they do. We will analyze the topic of sharing social media articles as we intend to learn more about the reasons why people spread certain information. Determining the driving forces that motivate people to share certain things such as articles can be imperative in funneling out the massive amounts of data in traffic of shared content through the Internet and social media and analyzing the important facets of an article or shareable content that would encourage someone to share it and allowing a digital company, such as Mashable, to discover factors that can perhaps boost their profit with more people sharing or reading their articles. If Mashable finds out factors that help the number of shares of their articles, the digitial company could provide articles that are strong in those factors so more people will read and share them, leading to greater success. To mine the data and determine what variables predict the amount of shares an article gets, data samples are taken from the digital media website Mashable with only the key variables of interest included in a spreadsheet. Given the information from Mashable articles gathered for over 2 years, we intend to answer the question of what variables can predict the amount of shares an article obtains.

Exploratory Data Analysis

After obtaining the spreadsheet of data about the social project, we will first familiarize ourselves with what is contained in the dataset. We can look at the first and last several rows of data regarding social media sharing and move onto univariate data visualization and analysis.

Observing Snippets of Data

It's important to note that the data comprises 4351 observations or articles with 5 variables. The days published variable is one of the 7 days in the week, sentiment being either positive or negative, channel being Business, Technology, Entertainment, World and Other, and content being categorized later in the analysis of categorical predictor variables. Regarding our variables of interest, we intend to predict the number of shares of a Mashable article from the variables of content, the day it was published, the overall positive or negative tone in the article, and the channel or type of website.

Variable	Description
shares	The number of article shares among sample of Mashable articles
content	The word count in the article
daypublished	The day during the week when the article was published
sentiment	The overall tone or sentiment, positivity or negativity, in the article
channel	The type of website or topic of the article (Business, Tech, (etc.))

To become familiar with the data, we will look at the first rows in the social project dataset:

```
## [1] "The First Six Rows of Mashable Article Dataset:"
##   shares content       channel daypublished sentiment
## 1   1500     745      Business       Monday  positive
## 2    727     342         Other       Monday  positive
## 3   2000     191          Tech     Thursday  positive
## 4    900     340 Entertainment    Wednesday  negative
## 5   3700     313         World      Tuesday  positive
## 6   1000     915 Entertainment    Wednesday  negative
```

```
## [1] "The Last Rows of Mashable Article Dataset"
```

```
##      shares content       channel daypublished sentiment
## 4346   5400     322         Other       Friday  negative
## 4347   2100     777      Business       Sunday  positive
## 4348   3100    1048      Business    Wednesday  positive
## 4349   1400    1617 Entertainment       Sunday  positive
## 4350   1600     580         Other     Thursday  positive
## 4351   1200    1373         World    Wednesday  positive
```

```
## [1] "Maximum Values For Number of Shares and Article Content"
```

```
##  shares content
##    8900    2365
```

```
## [1] "Minimum Values For Number of Shares and Article Content"
```

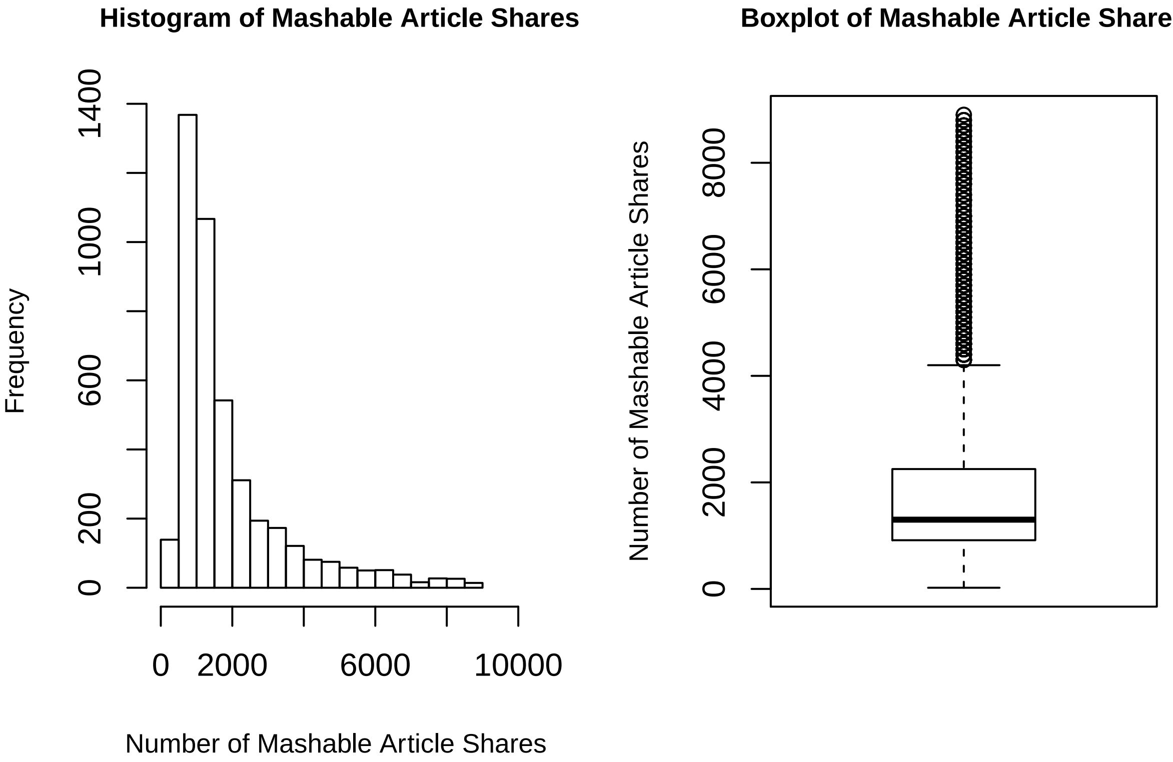
```
##  shares content
##      22       0
```

The magnitude of the shares are notably high and roughly in the high hundreds or thousands. The content, or word count, in the article appears to vary greatly from low hundreds all the way to the thousands. The sentiments and days published appear to be as expected where daypublished are the days of the week and the sentiment could either be defined as positive or negative. Notably, the maximum number of shares was 8900 shares with the lowest number of shares of an article being 22 shares. The largest word count was 2365 words and the lowest was 0 words. The other variables were categorical.

Univariate Exploratory Data Analysis

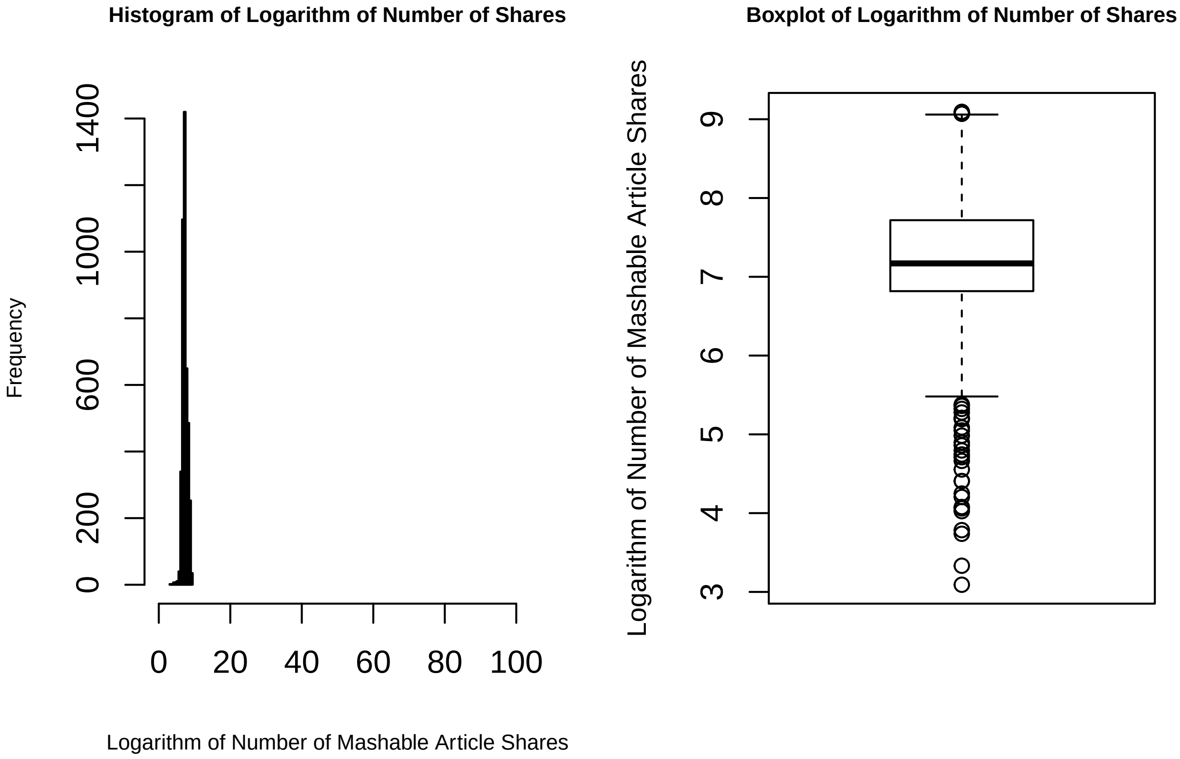
After observing the magnitude of the data variables, we will move onto observing the distribution of the response variable and explanatory variables in histograms, boxplots, barcharts as well as with summary statistics. It's important to note that for the ANOVA model that we will ultimately perform on the data, the content or word count in the articles has been categorized as $1 : 0 \ words \leq content < 200 \ words, \ 2 : 200 \ words \leq content < 400 \ words, \ 3 : 400 \ words \leq content < 600 \ words, \ 4 : 600 \ words \leq content < 800 \ words, \ 5 : 800 \ words \leq content < 1000 \ words, \ 6 : 1000 \ words \leq content < 1200 \ words, \ 7 : 1200 \ words \leq content < 1400 \ words, \ 8 : 1400 \ words \leq content < 1600 \ words, \ 9 : 1600 \ words \leq content < 1800 \ words, \ 10 : 1800 \ words \leq content < 2000 \ words, \ 11 : 2000 \ words \leq content < 2200 \ words, \ 12 : 2200 \ words \leq content < 2400 \ words.$ The content was divided into 4 categories because given that the largest value in the data set was 2365 and we were to categorize the word count by even numbers, 2400 can be divided by 12 to capture ranges of values in increments of 200 words. The choice for the categorization of content into 12 groups of 200 word increments was from the histogram of the quantity variable content, which yielded roughly 12 bins.

We will first observe the distribution of the quantitative response variable of the number of Mashable article shares.



```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       22     914    1300    1908    2250    8900
```

After constructing the histogram of the Mashable article shares, we observe a strongly right-skewed and unimodal distribution in the distribution of the number of Mashable article **shares**, shown in one mode within the distribution and most of the number of article shares being fairly low. There are many outliers in the number of shares as the number of data points beyond the upper fence of the boxplot, so we would have to transform the shares distribution. The median of number of article shares is at 1300 and the spread of the data is roughly 1336 shares, which is the range of shares that captures the middle 50% of the distribution of article shares. In the data visualizations below, we try transforming the distribution of shares by taking the logarithm of the number of article shares to improve the symmetry of the distribution of article shares.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.091   6.818   7.170   7.291   7.718   9.094
```

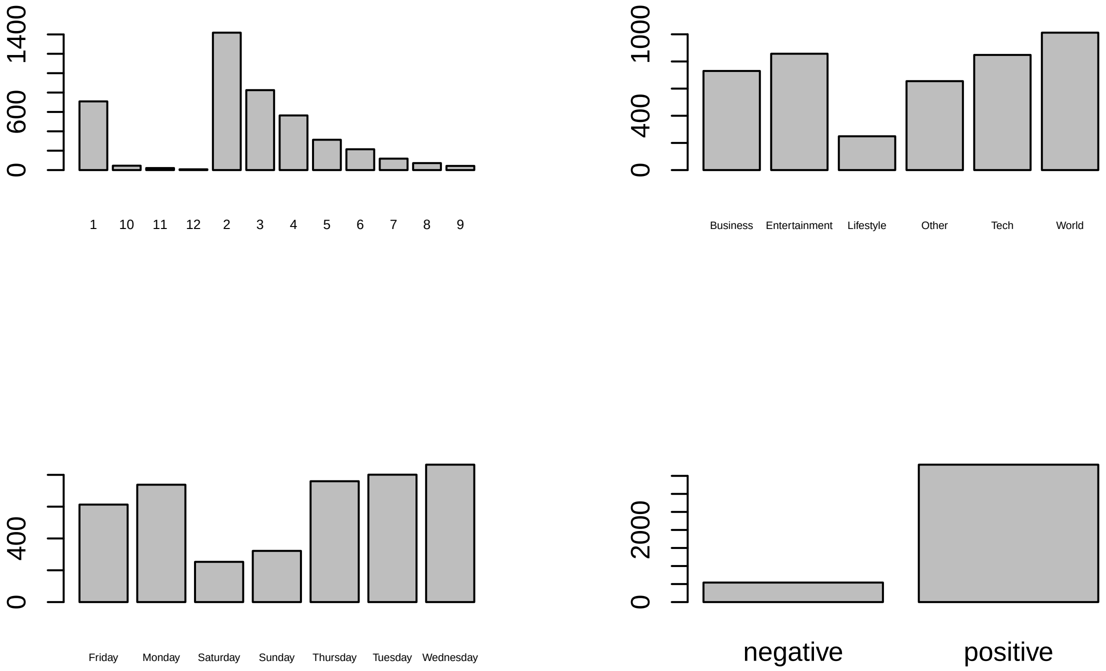
After having performing other transformations to improve the symmetry of the response variable and reduce outliers, the logarithmic transformation creates many small outliers although the histogram appears to be more normal and smaller fractional power transformations create both very small and very large outliers. The logarithmic transformation of the number of shares still yields outliers, but it's the most ideal transformation as the histogram appears to considerably more symmetric. The univariate exploratory data analysis reveals that the distribution of the logarithmic of the number of Mashable article shares is unimodal (one peak) and roughly symmetric, although still slightly right skewed from the boxplot There is still a considerable amount of outliers, but it could simply be an innate component of the data. When outliers are not a rarity, we would conclude the skewness and outliers in the data to be attributed to the data itself. The center measure is the median of the logarithm of the number of article shares is 7.170. The spread of the data can be observed in the range of the data (the minimum number of logarithm of shares is 3.091 and maximum logarithm of shares being 9.094) which is 6.003. The interquartile range of the logarithm of number of Mashable article shares is 0.9. We will continue with our exploratory data analysis by analyzing the categorical predictor variables in tabular and bar chart form.

```
## 
##    1   10   11   12    2    3    4    5    6    7    8    9 
##  709   45   21    8 1418  825  564  313  215  118   72   43
```

```
## 
##      Business Entertainment     Lifestyle         Other          Tech 
##           730           857           249           655           848 
##         World 
##          1012
```

```
## 
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday 
##       613       738       253       322       760       801       864
```

```
## 
## negative positive 
##      541     3810
```



The **categorical content, WordContent, variable** has 709 articles (16.30% of articles) that are between 0 and 200 words (200 exclusive), 1418 articles (32.59% of articles) between 200 and 400 words (400 exclusive), 825 articles (18.96% of articles) between 400 and 600 words (600 exclusive), 564 articles (12.96% of articles) between 600 words and 800 words (800 exclusive), 313 articles (7.19% of articles) between 800 and 1000 words (1000 exclusive), 215 articles (4.94% of all articles) between 1000 and 1200 words (1200 exclusive), 118 articles (2.71% of articles) between 1200 and 1400 words (1400 exclusive), 72 articles (1.65% of articles) between 1400 and 1600 words (1600 exclusive), 43 articles (0.988% of articles) between 1600 and 1800 words (1800 exclusive), 45 articles (1.03% of articles) between 1800 and 2000 words (2000 exclusive), 21 articles (0.48% of articles) between 2000 and 2200 words (2200 exclusive), and 8 articles (0.184% of articles) between 2200 and 2400 articles. Among the Mashable articles, a majority of them have a category 1 content with a word count that is between 0 and 600 words (600 exclusive). The category 12 content contains the fewest articles (8) having between 2200 and 2400 words (2400 exclusive). In general, aside from the transition from the first to second categories, for the content categories with ranges of higher word counts, there is a lower frequency of articles with that many words.

Among the Mashable articles, regarding their **channel, or type of website**, the channel containing the fewest articles was Lifestyle as 249 articles (5.72% of the articles) covered the topic of Lifestyle. 730 articles (16.78% of the articles) covered the Business channel, 655 articles (15.05% of the articles) covered the Other channel, 857 articles (19.70% of the articles) covered the Entertainment channel, 848 articles (19.49% of the articles) covered the Technology channel, and 1012 articles (23.26% of the articles) covered the World channel. The channel consisting of the most articles was the World channel with 1012 articles, 23.26% of all the articles.

The **daypublished, or the day when an article was published** had Saturday as the day with the fewest articles being published on as 253 articles (5.81% of articles) were published on Saturday. 322 articles (7.40% of the articles) were published on Sunday, 738 articles (16.96% of the articles) were published on Monday, 801 articles (18.41% of the articles) published on Tuesday, 864 articles (19.86% of the articles) published on Wednesday, 760 articles (17.47% of the articles) published on Thursday, and 613 articles (14.09% of the articles) published on Friday. Most of the articles were published on Wednesday (19.86%, nearly 20% of the articles (864 articles) were published on Wednesday).

The **sentiment** among the articles had an overwhelming majority of positive sentiment. There were 3810 articles (87.57% of the articles) that constituted positive sentiment and 541 articles (12.43% of the articles) with negative sentiment. Many more articles were positive than negative.

Bivariate Exploratory Data Analysis

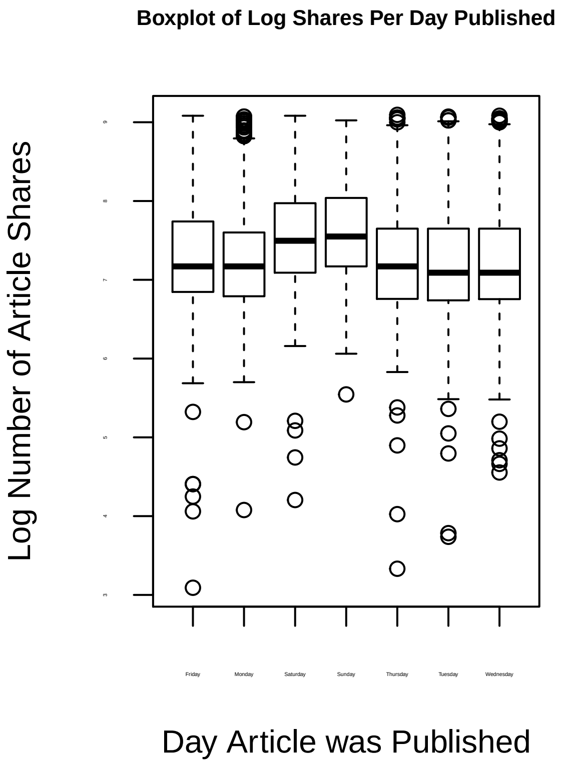
In terms of bivariate data analysis, we will observe the boxplots of each predictor against the response variable of the number of shares Mashable articles got. We will then produce interaction plots between each combination of the two predictors to determine if there's any notable interaction between the categorical predictors.

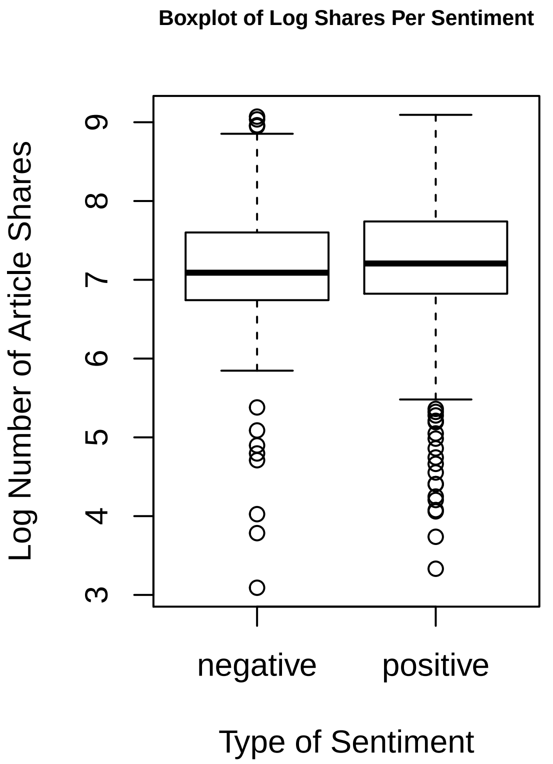
```
## # A tibble: 7 x 4
##   daypublished `mean(log(shares))` `sd(log(shares))` `n()`
##   <fct>                      <dbl>             <dbl> <int>
## 1 Friday                      7.29             0.707   613
## 2 Monday                      7.24             0.682   738
## 3 Saturday                    7.57             0.718   253
## 4 Sunday                      7.62             0.633   322
## 5 Thursday                    7.24             0.711   760
## 6 Tuesday                     7.23             0.727   801
## 7 Wednesday                   7.23             0.700   864
```

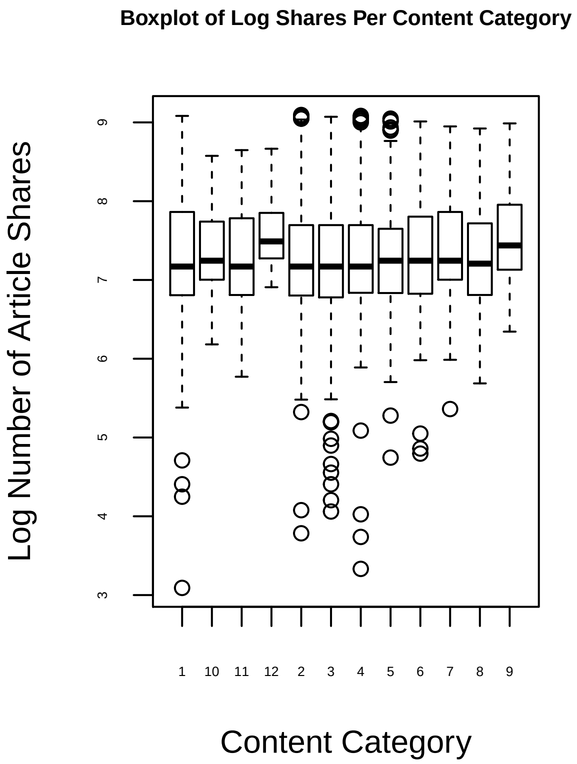
```
## # A tibble: 2 x 4
##   sentiment `mean(log(shares))` `sd(log(shares))` `n()`
##   <fct>                   <dbl>             <dbl> <int>
## 1 negative                 7.19             0.742   541
## 2 positive                 7.31             0.706  3810
```

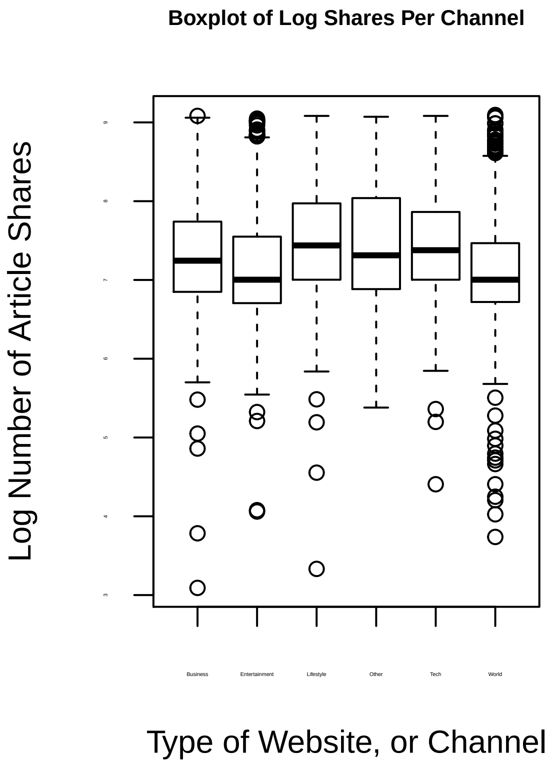
```
## # A tibble: 12 x 4
##    WordContent `mean(log(shares))` `sd(log(shares))` `n()`
##    <chr>                     <dbl>             <dbl> <int>
##  1 1                          7.31             0.750   709
##  2 10                         7.37             0.591    45
##  3 11                         7.26             0.766    21
##  4 12                         7.60             0.544     8
##  5 2                          7.28             0.695  1418
##  6 3                          7.25             0.733   825
##  7 4                          7.29             0.706   564
##  8 5                          7.28             0.662   313
##  9 6                          7.33             0.734   215
## 10 7                          7.40             0.666   118
## 11 8                          7.32             0.742    72
## 12 9                          7.54             0.609    43
```

```
## # A tibble: 6 x 4
##   channel       `mean(log(shares))` `sd(log(shares))` `n()`
##   <fct>                       <dbl>             <dbl> <int>
## 1 Business                     7.31             0.713   730
## 2 Entertainment                7.17             0.701   857
## 3 Lifestyle                    7.47             0.763   249
## 4 Other                        7.44             0.739   655
## 5 Tech                         7.45             0.655   848
## 6 World                        7.10             0.670  1012
```

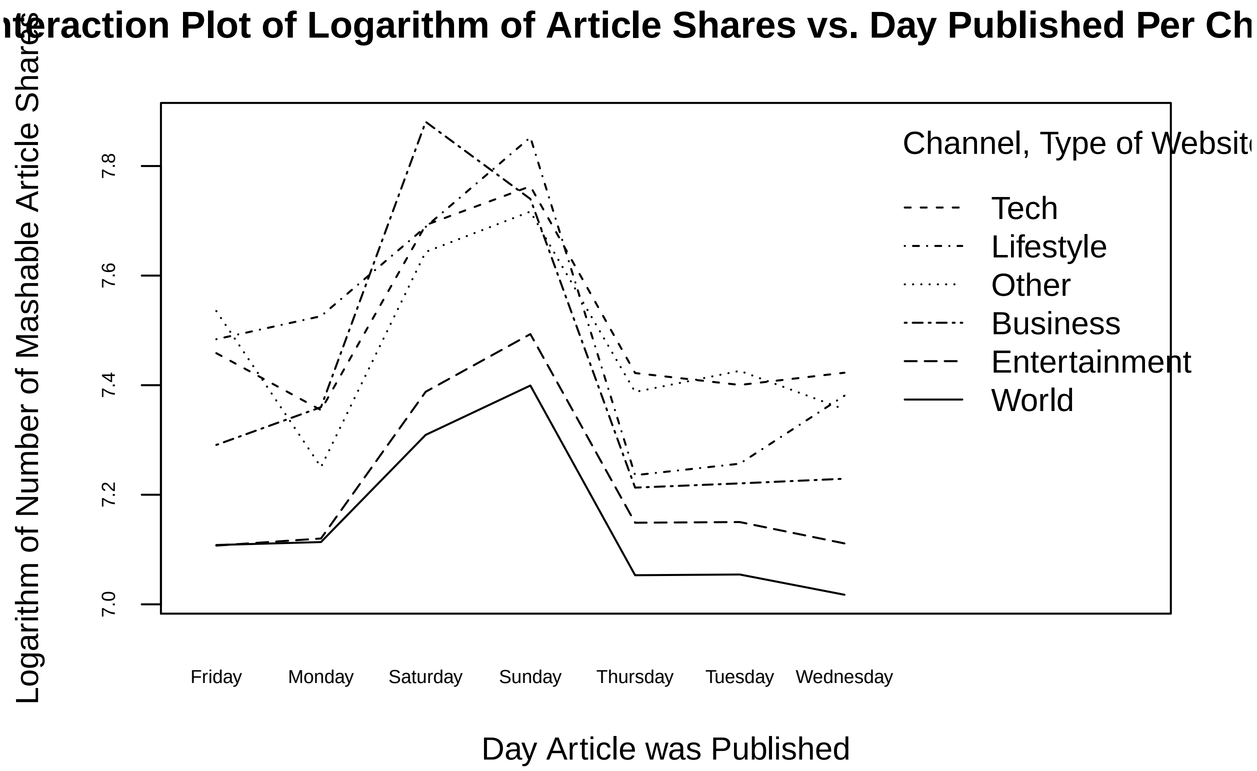

Boxplot of Log Shares Per Day Published
Log Number of Article Shares
3
4
5
6
7
8
9
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day Article was Published

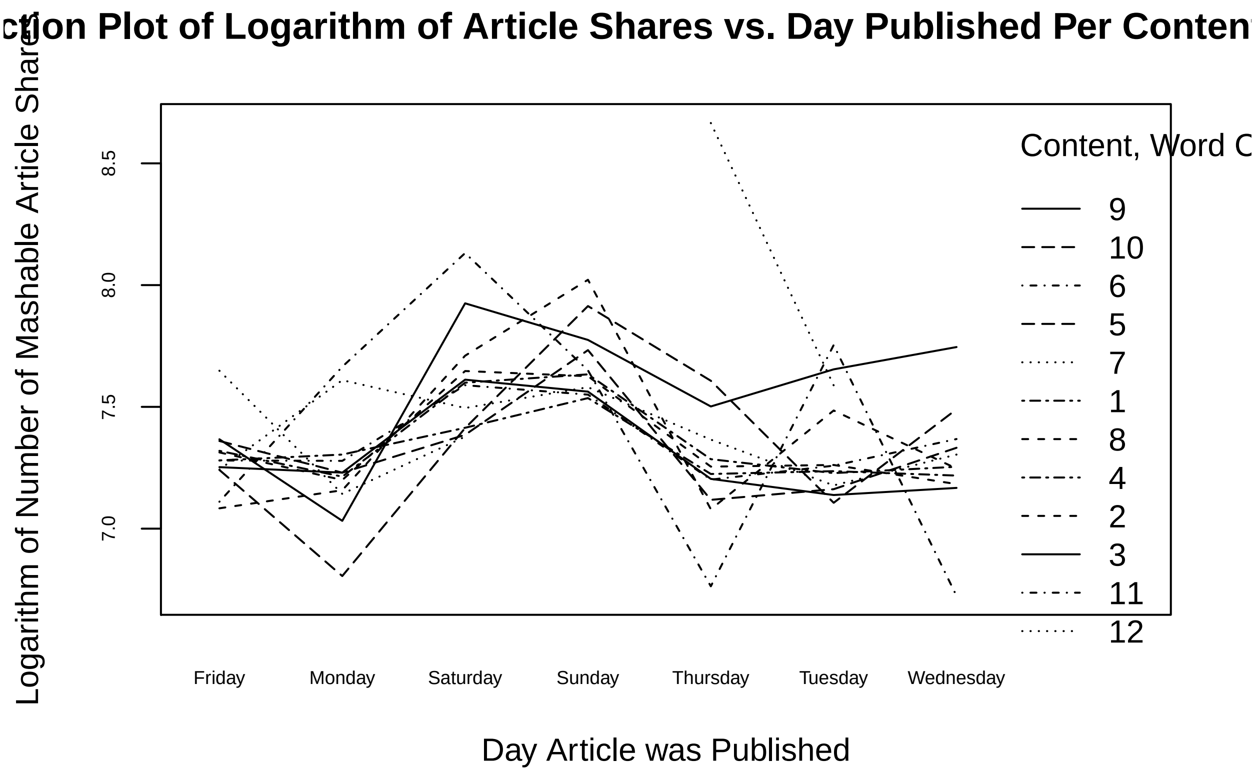

Boxplot of Log Shares Per Sentiment
Log Number of Article Shares
3
4
5
6
7
8
9
negative
positive
Type of Sentiment

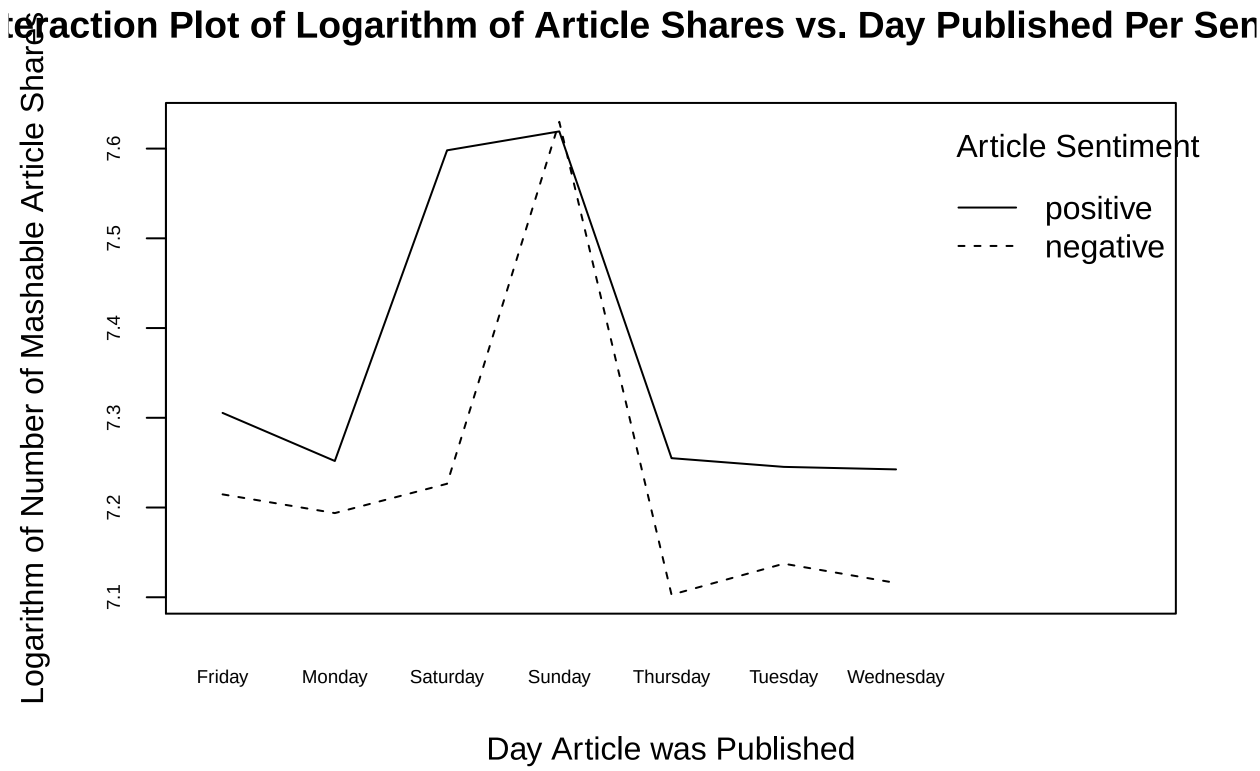

Boxplot of Log Shares Per Content Category
Log Number of Article Shares
3
4
5
6
7
8
9
1
10
11
12
2
3
4
5
6
7
8
9
Content Category

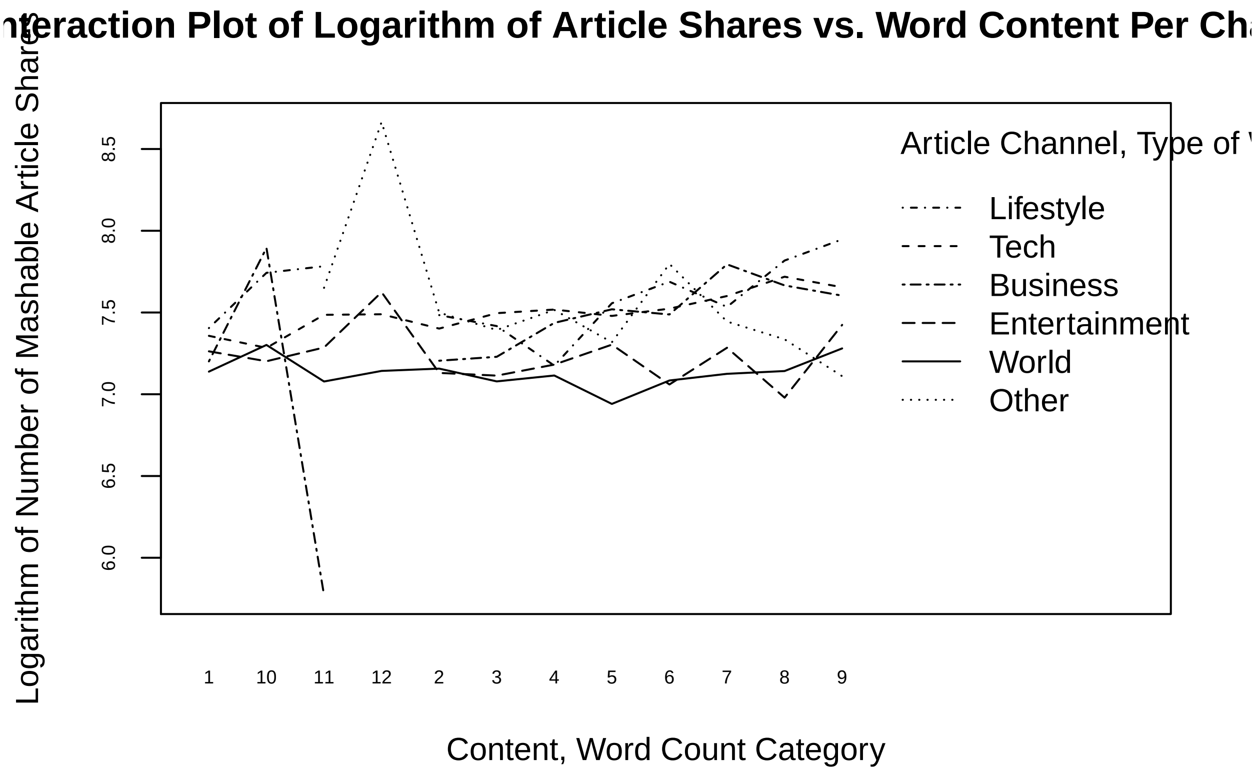

Boxplot of Log Shares Per Channel
Log Number of Article Shares
3
4
5
6
7
8
9
Business
Entertainment
Lifestyle
Other
Tech
World
Type of Website, or Channel

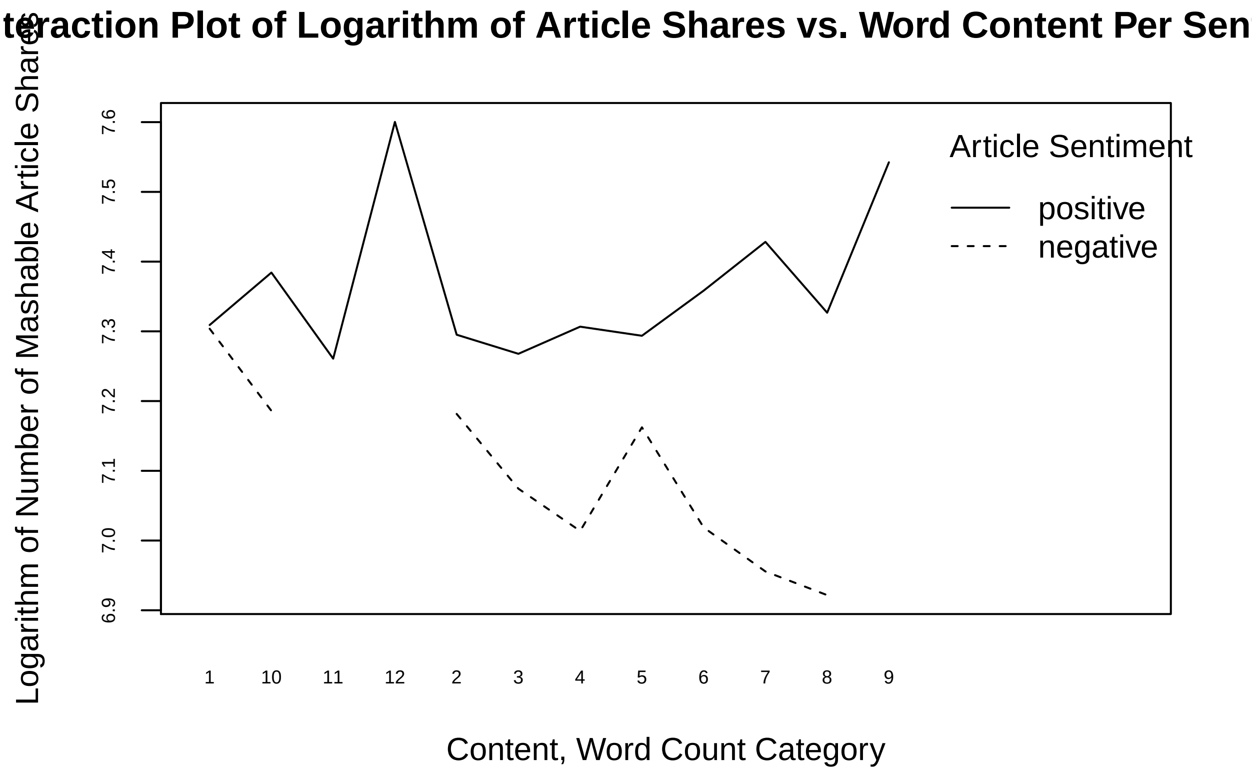
From summary tables of the side-by-side boxplots and the plots itself for each categorical predictor variable, we observe some relationships. The boxplots all contain outliers, but that remains a characteristic of the number of article share data, after transformed by logarithm. In the boxplots of the response variable of the logarithm of number of article shares plotted against each categorical predictor, every plot contains significant outliers, but that likely has something to do with the The boxplots for the logarithm of number of article shares against the type of sentiment of the article reveal that positive articles have slightly greater value of logarithm of article shares, on average, than negative articles. The spread, standard deviation, in the logarithm of article shares is essentially the same between positive and negative sentiment (0.706 vs. 0.742) although negative sentiment has slightly greater variability. In the boxplots for categorized content, there is a slight, but insignificant difference in logarithm of article shares between the content categories (or word count categories) in the articles as the medians all appear to be essentially the same except for slightly higher medians for category group 9 (1600 to 1800 words) and group 12 (2200 to 2400 words). The category with the greatest spread, 0.742, is 7 (1200 to 1400 words) while category 9 (1600 to 1800 words) has the least variability, 0.591. The boxplots for the day an article was published appears to convey a significant difference as the logarithm of the number of article shares appeared to be higher on average over the weekend than the weekdays. The day with the most variability is on Tuesday (0.727) while Sunday seems to have the least variability (0.633). Also, there appears to be noticeable differences in the number of article shares between the channels in their boxplots. Namely, articles from the Lifestyle channel appear to be shared the most (7.47) while the articles from Entertainment (7.17) and World (7.10) channels seem to be shared the least. The Tech channel appears to have the least variability (0.655) while Lifestyle channel had the highest variability (0.763).

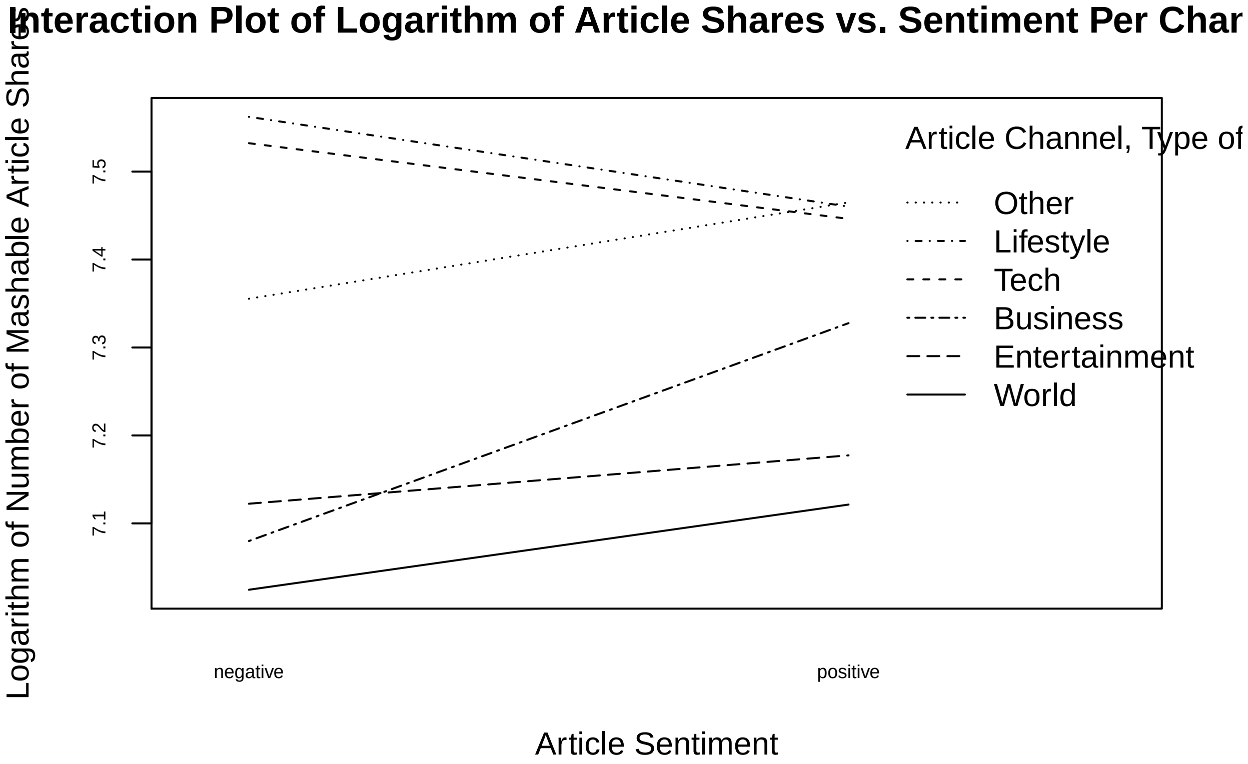

nteraction Plot of Logarithm of Article Shares vs. Day Published Per Ch
Logarithm of Number of Mashable Article Share
7.0
7.2
7.4
7.6
7.8
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day Article was Published
Channel, Type of Websit
Tech
Lifestyle
Other
Business
Entertainment
World


ction Plot of Logarithm of Article Shares vs. Day Published Per Conten
Logarithm of Number of Mashable Article Share
7.0
7.5
8.0
8.5
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day Article was Published
Content, Word C
9
10
6
5
7
1
8
4
2
3
11
12


teraction Plot of Logarithm of Article Shares vs. Day Published Per Sen
Logarithm of Number of Mashable Article Shares
7.1
7.2
7.3
7.4
7.5
7.6
Article Sentiment
positive
negative
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day Article was Published


nteraction Plot of Logarithm of Article Shares vs. Word Content Per Ch
Logarithm of Number of Mashable Article Share
6.0
6.5
7.0
7.5
8.0
8.5
1
10
11
12
2
3
4
5
6
7
8
9
Content, Word Count Category
Article Channel, Type of
Lifestyle
Tech
Business
Entertainment
World
Other


teraction Plot of Logarithm of Article Shares vs. Word Content Per Sen
Logarithm of Number of Mashable Article Shares
6.9
7.0
7.1
7.2
7.3
7.4
7.5
7.6
Article Sentiment
positive
negative
1
10
11
12
2
3
4
5
6
7
8
9
Content, Word Count Category



These interaction plots collectively seem to explain individual main effects for the day an article was published, the channel, or type of website, and sentiment, but most often don't reveal interaction effects as there isn't apparent change in trace factor effect over different levels of the x factor. However, the interaction of the logarithm of articles shares against the word count, or content, categorized per channel (type of website) seems to reveal an interaction effect as Business channel appears to have a different relationship over content categorized than, say the other channel. We can determine the specific significant interactions between each combination of dummy variables for each categorical predictor of by, after having determined interaction terms that are significant, running models containing those interaction terms of specific significant dummy variables. This will be conducted in the modeling section.

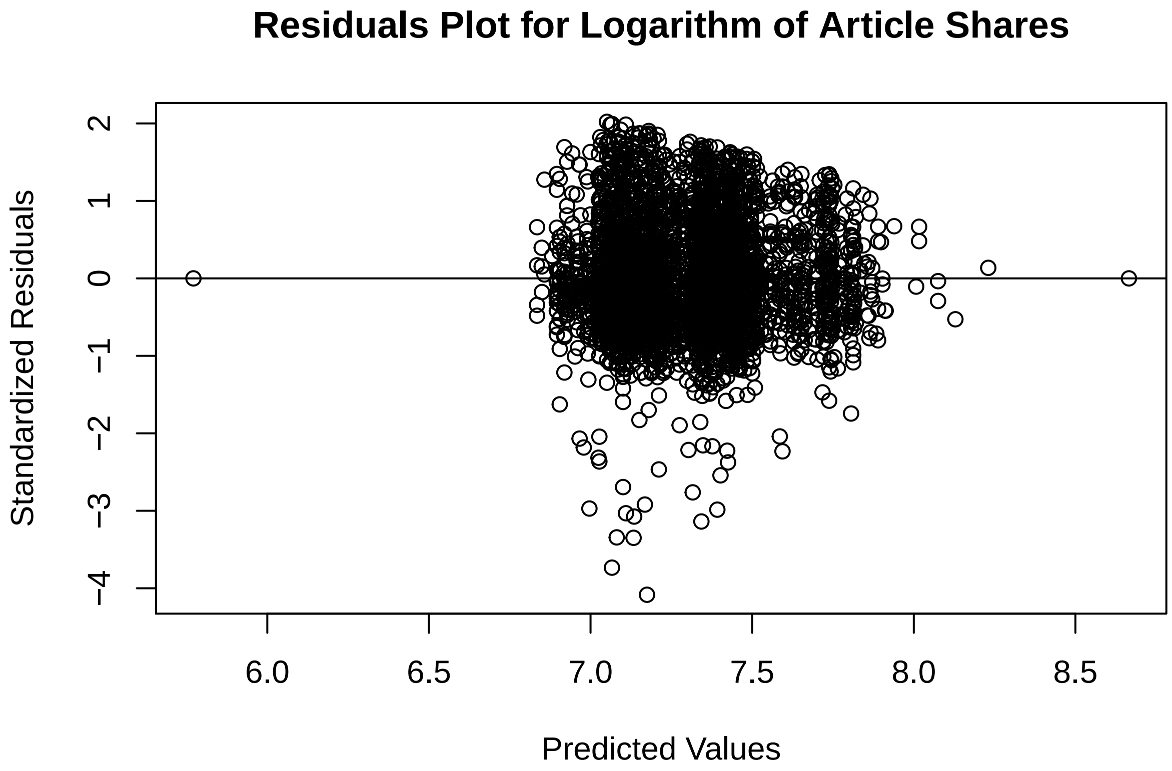
Modeling

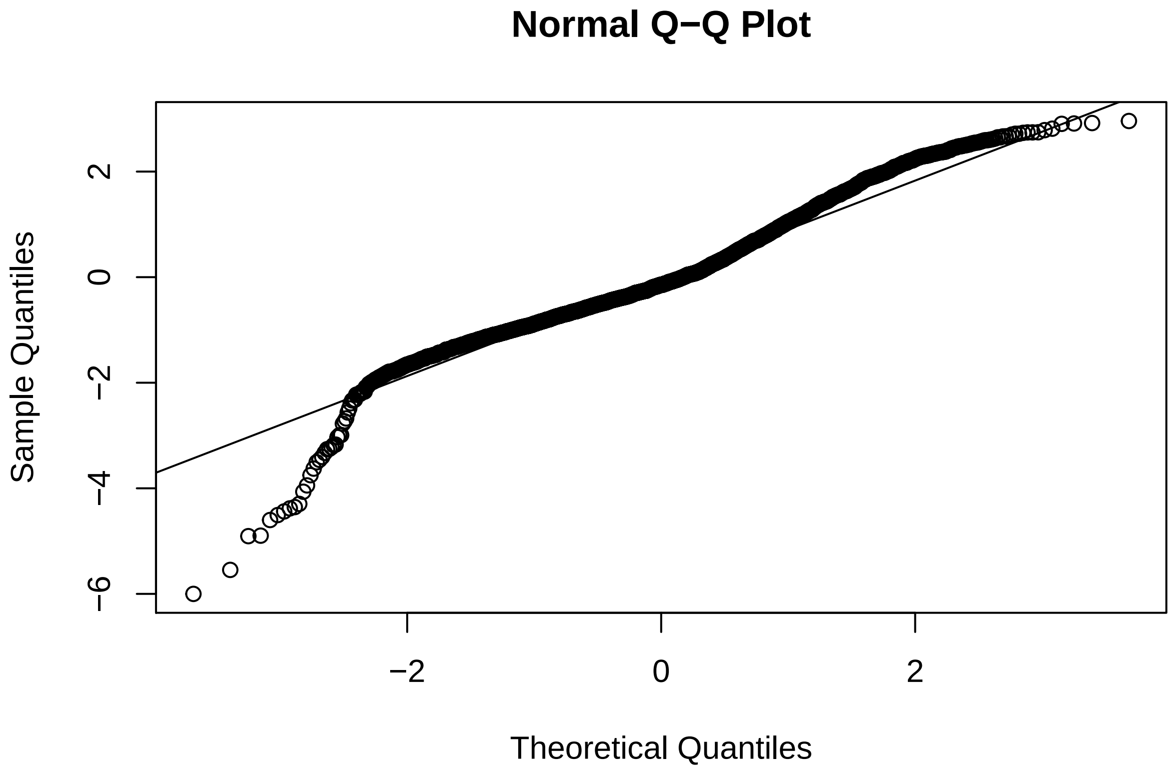
We will construct an ANOVA model in order to predict article shares from the explanatory variables. Given that there are 4 factors in this ANOVA, we must consider creating a factorial ANOVA model with interaction terms. We acknowledge that there could be interactions be three categorical predictors, but most often they result in very insignificant terms. We run a model of the ANOVA with the possible interaction terms between the categorical predictors, then rerun the model with the significant interaction term as a significant interaction term would suffice for a 4 factor ANOVA interaction model. Also, as aforementioned after transforming the response variable by a logarithm, performing other trasnformations such as square root or other fractional powers didn't improve the outliers or the normality of the response variable. The logarithm transformation was the best option for the model.

```
##                                            Df Sum Sq Mean Sq F value
## factor(WordContent)                        11    7.4   0.676   1.435
## factor(channel)                             5   94.6  18.929  40.201
## factor(daypublished)                        6   58.8   9.795  20.803
## factor(sentiment)                           1    1.9   1.947   4.136
## factor(WordContent):factor(channel)        52   34.2   0.658   1.397
## factor(WordContent):factor(daypublished)   62   25.0   0.404   0.858
## factor(WordContent):factor(sentiment)       8    2.5   0.307   0.653
## factor(channel):factor(daypublished)       30   11.5   0.383   0.814
## factor(channel):factor(sentiment)           5    3.0   0.591   1.255
## factor(daypublished):factor(sentiment)      6    2.4   0.401   0.852
## Residuals                                4164 1960.6   0.471
##                                          Pr(>F)
## factor(WordContent)                       0.150
## factor(channel)                          <2e-16 ***
## factor(daypublished)                     <2e-16 ***
## factor(sentiment)                         0.042 *
## factor(WordContent):factor(channel)       0.032 *
## factor(WordContent):factor(daypublished)  0.778
## factor(WordContent):factor(sentiment)     0.734
## factor(channel):factor(daypublished)      0.752
## factor(channel):factor(sentiment)         0.280
## factor(daypublished):factor(sentiment)    0.530
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We notice that the model run with the possible pair interaction terms results in highly significant terms (p-values far below 0.05 at <10^-16) of the channel and daypublished of the article with sentiment also being significant at the 5% significance level (0.0331). The interaction term between content (word count) categorized and the channel of the article (p-value below 0.05 at 0.0173) being significant at the 5% significance level. However, the categorized word content is not extremely significant in predicting the number of logarithm of article shares, which is likely attributed to its highly right-skewed distribution as most articles don't have fairly high word count.

```
##                                       Df Sum Sq Mean Sq F value Pr(>F)
## factor(WordContent)                   11    7.4   0.676   1.441 0.1473
## factor(channel)                        5   94.6  18.929  40.360 <2e-16 ***
## factor(daypublished)                   6   58.8   9.795  20.885 <2e-16 ***
## factor(sentiment)                      1    1.9   1.947   4.152 0.0416 *
## factor(WordContent):factor(channel)   52   34.2   0.658   1.402 0.0304 *
## Residuals                           4275 2005.0   0.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals Plot for Logarithm of Article Shares
Standardized Residuals
2
1
0
−1
−2
−3
−4
6.0
6.5
7.0
7.5
8.0
8.5
Predicted Values


Normal Q−Q Plot
Sample Quantiles
2
0
−2
−4
−6
−2
0
2
Theoretical Quantiles

Having performed other transformations that didn't create a symmetric enough distribution of the response variable (other fractional powers), the logarithmic response variable was the best transformation. Also, if the significant interaction term was excluded from the model, the normal probability plot would deviate significantly more than in the final model with the interaction term included. We would assume that the individual groups (combination of the levels of each categorical predictor) within the data is independent of each other. From the residuals plot of the final social media ANOVA model containing the significant interaction between the article content and channel, type of website, and the other four factors with a logarithm response variable, the residuals appear to have a mean of roughly 0 throughout the residuals plot as an equal concentration of data points are below the 0 residual line and above it, creating roughly a mean of 0, decently satisfying the error mean of 0 condition. The errors appear to be independent as they don't follow a trend and don't seem to be associated in a relationship, so the independent errors condition seems to be satisfied. The condition of there being equal variance, fixed standard deviation, in the residuals plot also appears to be decently satsfied as the width of the residual plot data points remains the same throughout the plot. However, after filtering for the significant interaction terms and implementing a logarithm transformation to the response variable to improve symmetry and normality, the normal probability plot isn't normal for lower quartiles and very high quartiles as data deviates more from the qqline (more for lower quartiles and less for higher quartiles) - the qqline being where sample quantiles match the normal distribution quantiles. This perhaps is attributed to the nature of the number of shares data being highly right skewed, so it may be difficult to improve normality if the data is innately right skewed. Regardless, after transforming the response variable and including appropriate interaction terms, the social media model appears to be the best result after considering remedies to improve the conditions for the ANOVA test and the normality in the normal probability plot is in fact very well as most of the qqplot aligns with the normal quantiles in the qqline.

```
##
## Call:
## lm(formula = log(shares) ~ WordContent + channel + daypublished +
##     sentiment + WordContent:channel, data = social)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0837 -0.4380 -0.0981  0.4054  2.0204
##
## Coefficients: (3 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         7.1747179  0.0764987  93.789  < 2e-16
## WordContent10                       0.7181271  0.2874321   2.498  0.01251
## WordContent11                      -1.4000987  0.6882731  -2.034  0.04199
## WordContent12                       0.0162154  0.4915450   0.033  0.97369
## WordContent2                        0.0083900  0.0782353   0.107  0.91460
## WordContent3                        0.0124840  0.0908243   0.137  0.89068
## WordContent4                        0.1886665  0.0949928   1.986  0.04708
## WordContent5                        0.2608921  0.1128353   2.312  0.02082
## WordContent6                        0.2305346  0.1171223   1.968  0.04910
## WordContent7                        0.5201926  0.1946693   2.672  0.00756
## WordContent8                        0.3559704  0.2512144   1.417  0.15656
## WordContent9                        0.3291918  0.2510311   1.311  0.18981
## channelEntertainment                0.0304969  0.0898641   0.339  0.73435
## channelLifestyle                    0.1541102  0.1743452   0.884  0.37678
## channelOther                        0.1653553  0.0787799   2.099  0.03588
## channelTech                         0.1473374  0.0908289   1.622  0.10485
## channelWorld                       -0.0767443  0.1033561  -0.743  0.45781
## daypublishedMonday                 -0.0416181  0.0378121  -1.101  0.27111
## daypublishedSaturday                0.2422461  0.0518155   4.675 3.03e-06
## daypublishedSunday                  0.3100302  0.0475919   6.514 8.15e-11
## daypublishedThursday              -0.0644192  0.0374790  -1.719  0.08572
## daypublishedTuesday               -0.0499956  0.0370530  -1.349  0.17731
## daypublishedWednesday             -0.0733644  0.0364484  -2.013  0.04420
## sentimentpositive                  0.0701863  0.0331473   2.117  0.03428
## WordContent10:channelEntertainment -0.7930391  0.3295665  -2.406  0.01616
## WordContent11:channelEntertainment  1.4548410  0.7323585   1.987  0.04704
## WordContent12:channelEntertainment  0.3659471  0.6338881   0.577  0.56376
## WordContent2:channelEntertainment  -0.1326358  0.1077915  -1.230  0.21858
## WordContent3:channelEntertainment  -0.1529383  0.1226557  -1.247  0.21251
## WordContent4:channelEntertainment  -0.2677853  0.1319097  -2.030  0.04241
## WordContent5:channelEntertainment  -0.2403469  0.1625354  -1.479  0.13928
## WordContent6:channelEntertainment  -0.4329461  0.1656274  -2.614  0.00898
## WordContent7:channelEntertainment  -0.5244628  0.2328652  -2.252  0.02436
## WordContent8:channelEntertainment  -0.6380955  0.2914256  -2.190  0.02861
## WordContent9:channelEntertainment  -0.2112516  0.3206425  -0.659  0.51003
## WordContent10:channelLifestyle     -0.3273982  0.4753958  -0.689  0.49106
## WordContent11:channelLifestyle      1.8343039  0.9846837   1.863  0.06255
## WordContent12:channelLifestyle             NA         NA      NA       NA
## WordContent2:channelLifestyle       0.0762214  0.1985420   0.384  0.70107
## WordContent3:channelLifestyle      -0.0219328  0.2030311  -0.108  0.91398
## WordContent4:channelLifestyle      -0.4569711  0.2240972  -2.039  0.04149
## WordContent5:channelLifestyle      -0.1038656  0.2384853  -0.436  0.66321
## WordContent6:channelLifestyle      -0.0001992  0.2755355  -0.001  0.99942
## WordContent7:channelLifestyle      -0.4237302  0.3264770  -1.298  0.19440
## WordContent8:channelLifestyle       0.1076580  0.3845866   0.280  0.77954
## WordContent9:channelLifestyle       0.1922504  0.4544970   0.423  0.67232
## WordContent10:channelOther                 NA         NA      NA       NA
## WordContent11:channelOther          1.5445391  0.8430613   1.832  0.06701
## WordContent12:channelOther          1.3035575  0.8442911   1.544  0.12267
## WordContent2:channelOther           0.0841205  0.1008763   0.834  0.40439
## WordContent3:channelOther          -0.0353871  0.1244122  -0.284  0.77609
## WordContent4:channelOther          -0.0426793  0.1505256  -0.284  0.77678
## WordContent5:channelOther          -0.3705947  0.1895162  -1.955  0.05059
## WordContent6:channelOther           0.1779111  0.2339037   0.761  0.44693
## WordContent7:channelOther          -0.5219513  0.3963030  -1.317  0.18789
## WordContent8:channelOther          -0.4658389  0.3633214  -1.282  0.19985
## WordContent9:channelOther          -0.5967539  0.5475500  -1.090  0.27584
## WordContent10:channelTech          -0.8415070  0.4060322  -2.073  0.03828
## WordContent11:channelTech           1.4631682  0.7713250   1.897  0.05790
## WordContent12:channelTech          -0.0402493  0.6934627  -0.058  0.95372
## WordContent2:channelTech            0.0195533  0.1076238   0.182  0.85584
## WordContent3:channelTech            0.0914589  0.1231523   0.743  0.45773
## WordContent4:channelTech           -0.0849238  0.1330545  -0.638  0.52334
## WordContent5:channelTech           -0.1595821  0.1551853  -1.028  0.30385
## WordContent6:channelTech           -0.0725825  0.1732143  -0.419  0.67521
## WordContent7:channelTech           -0.2685540  0.2461850  -1.091  0.27539
## WordContent8:channelTech           -0.0417582  0.3375643  -0.124  0.90156
## WordContent9:channelTech           -0.1283153  0.3256155  -0.394  0.69355
## WordContent10:channelWorld         -0.5877780  0.3842260  -1.530  0.12615
## WordContent11:channelWorld          1.3651086  0.7574876   1.802  0.07159
## WordContent12:channelWorld                 NA         NA      NA       NA
## WordContent2:channelWorld          -0.0027055  0.1185015  -0.023  0.98179
## WordContent3:channelWorld          -0.0799848  0.1285298  -0.622  0.53377
## WordContent4:channelWorld          -0.2258301  0.1331827  -1.696  0.09003
## WordContent5:channelWorld          -0.4599208  0.1541775  -2.983  0.00287
## WordContent6:channelWorld          -0.2996168  0.1689290  -1.774  0.07620
## WordContent7:channelWorld          -0.5751504  0.2481016  -2.318  0.02049
## WordContent8:channelWorld          -0.4304562  0.3248394  -1.325  0.18520
## WordContent9:channelWorld          -0.1884463  0.4322777  -0.436  0.66290
##
## (Intercept)                        ***
## WordContent10                      *
## WordContent11                      *
## WordContent12
## WordContent2
## WordContent3
## WordContent4                       *
## WordContent5                       *
## WordContent6                       *
## WordContent7                       **
## WordContent8
## WordContent9
## channelEntertainment
## channelLifestyle
## channelOther                       *
## channelTech
## channelWorld
## daypublishedMonday
## daypublishedSaturday               ***
## daypublishedSunday                 ***
## daypublishedThursday               .
## daypublishedTuesday
## daypublishedWednesday              *
## sentimentpositive                  *
## WordContent10:channelEntertainment *
## WordContent11:channelEntertainment *
## WordContent12:channelEntertainment
## WordContent2:channelEntertainment
## WordContent3:channelEntertainment
## WordContent4:channelEntertainment  *
## WordContent5:channelEntertainment
## WordContent6:channelEntertainment  **
## WordContent7:channelEntertainment  *
## WordContent8:channelEntertainment  *
## WordContent9:channelEntertainment
## WordContent10:channelLifestyle
## WordContent11:channelLifestyle     .
## WordContent12:channelLifestyle
## WordContent2:channelLifestyle
## WordContent3:channelLifestyle
## WordContent4:channelLifestyle      *
## WordContent5:channelLifestyle
## WordContent6:channelLifestyle
## WordContent7:channelLifestyle
## WordContent8:channelLifestyle
## WordContent9:channelLifestyle
## WordContent10:channelOther
## WordContent11:channelOther         .
## WordContent12:channelOther
## WordContent2:channelOther
## WordContent3:channelOther
## WordContent4:channelOther
## WordContent5:channelOther          .
## WordContent6:channelOther
## WordContent7:channelOther
## WordContent8:channelOther
## WordContent9:channelOther
## WordContent10:channelTech          *
## WordContent11:channelTech          .
## WordContent12:channelTech
## WordContent2:channelTech
## WordContent3:channelTech
## WordContent4:channelTech
## WordContent5:channelTech
## WordContent6:channelTech
## WordContent7:channelTech
## WordContent8:channelTech
## WordContent9:channelTech
## WordContent10:channelWorld
## WordContent11:channelWorld         .
## WordContent12:channelWorld
## WordContent2:channelWorld
## WordContent3:channelWorld
## WordContent4:channelWorld          .
## WordContent5:channelWorld          **
## WordContent6:channelWorld          .
## WordContent7:channelWorld          *
## WordContent8:channelWorld
## WordContent9:channelWorld
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6848 on 4275 degrees of freedom
## Multiple R-squared:  0.08946,    Adjusted R-squared:  0.07349
## F-statistic:   5.6 on 75 and 4275 DF,  p-value: < 2.2e-16
```

After running the multiple linear regression containing the interaction term of interest between the word count (WordContent) and channel of the article, we can deduce that the interaction between categorical predictors of WordCount and channel doesn't occur at every dummy variable level. The significant interactions occur between the WordContent category of 10 and channel of Entertainment, WordContent category of 5 and channel of Entertainment, WordContent category of 6 and channel of Entertainment, WordContent category of 7 and channel of Entertainment (at the 10% significance level), WordContent category of 9 and channel of Entertainment, WordContent category of 10 and channel of Lifestyle (significant at 10% level), WordContent category of 3 and channel of Lifestyle, WordContent category of 6 and channel of Lifestyle (significant at 10% level), WordContent category of 10 and channel of Other, WordContent of 4 and channel of Other, WordContent category of 9 and channel of Tech, WordContent category of 10 and channel of Tech (at the 10% significance level), WordContent category of 3,4, 5 and 6 with channel of World, and WordContent category of 10 and channel of World (significant at 10% level).

```
##                                           (Intercept)
##                                          7.1747178902
##                                 factor(WordContent)10
##                                          0.7181270739
##                              factor(WordContent)11
##                                      -1.4000986877
##                              factor(WordContent)12
##                                       0.0162154001
##                               factor(WordContent)2
##                                       0.0083900368
##                               factor(WordContent)3
##                                       0.0124840328
##                               factor(WordContent)4
##                                       0.1886665188
##                               factor(WordContent)5
##                                       0.2608920787
##                               factor(WordContent)6
##                                       0.2305345531
##                               factor(WordContent)7
##                                       0.5201925848
##                               factor(WordContent)8
##                                       0.3559704406
##                               factor(WordContent)9
##                                       0.3291918361
##                       factor(channel)Entertainment
##                                       0.0304969443
##                           factor(channel)Lifestyle
##                                       0.1541102340
##                               factor(channel)Other
##                                       0.1653553291
##                                factor(channel)Tech
##                                       0.1473373924
##                               factor(channel)World
##                                      -0.0767443480
##                         factor(daypublished)Monday
##                                      -0.0416181227
##                       factor(daypublished)Saturday
##                                       0.2422460863
##                         factor(daypublished)Sunday
##                                       0.3100302241
##                       factor(daypublished)Thursday
##                                      -0.0644191945
##                        factor(daypublished)Tuesday
##                                      -0.0499955891
##                      factor(daypublished)Wednesday
##                                      -0.0733643534
##                          factor(sentiment)positive
##                                       0.0701862740
## factor(WordContent)10:factor(channel)Entertainment
##                                      -0.7930390766
## factor(WordContent)11:factor(channel)Entertainment
##                                       1.4548410305
## factor(WordContent)12:factor(channel)Entertainment
##                                       0.3659471343
##  factor(WordContent)2:factor(channel)Entertainment
##                                      -0.1326358398
##  factor(WordContent)3:factor(channel)Entertainment
##                                      -0.1529383428
##  factor(WordContent)4:factor(channel)Entertainment
##                                      -0.2677853238
##  factor(WordContent)5:factor(channel)Entertainment
##                                      -0.2403469084
##  factor(WordContent)6:factor(channel)Entertainment
##                                      -0.4329460996
##  factor(WordContent)7:factor(channel)Entertainment
##                                      -0.5244628328
##  factor(WordContent)8:factor(channel)Entertainment
##                                      -0.6380955215
##  factor(WordContent)9:factor(channel)Entertainment
##                                      -0.2112516304
##     factor(WordContent)10:factor(channel)Lifestyle
##                                      -0.3273982390
##     factor(WordContent)11:factor(channel)Lifestyle
##                                       1.8343038949
##      factor(WordContent)2:factor(channel)Lifestyle
##                                       0.0762214150
##      factor(WordContent)3:factor(channel)Lifestyle
##                                      -0.0219327744
##      factor(WordContent)4:factor(channel)Lifestyle
##                                      -0.4569711043
##      factor(WordContent)5:factor(channel)Lifestyle
##                                      -0.1038655652
##      factor(WordContent)6:factor(channel)Lifestyle
##                                      -0.0001991687
##      factor(WordContent)7:factor(channel)Lifestyle
##                                      -0.4237302364
##      factor(WordContent)8:factor(channel)Lifestyle
##                                       0.1076579972
##      factor(WordContent)9:factor(channel)Lifestyle
##                                       0.1922504373
##         factor(WordContent)11:factor(channel)Other
##                                       1.5445390728
##         factor(WordContent)12:factor(channel)Other
##                                       1.3035574976
##          factor(WordContent)2:factor(channel)Other
##                                       0.0841204843
##          factor(WordContent)3:factor(channel)Other
##                                      -0.0353871038
##          factor(WordContent)4:factor(channel)Other
##                                      -0.0426792754
##          factor(WordContent)5:factor(channel)Other
##                                      -0.3705946985
##          factor(WordContent)6:factor(channel)Other
##                                       0.1779111340
##          factor(WordContent)7:factor(channel)Other
##                                      -0.5219512556
##          factor(WordContent)8:factor(channel)Other
##                                      -0.4658389103
##          factor(WordContent)9:factor(channel)Other
##                                      -0.5967538991
##          factor(WordContent)10:factor(channel)Tech
##                                      -0.8415070024
##          factor(WordContent)11:factor(channel)Tech 
##                                       1.4631681800 
##          factor(WordContent)12:factor(channel)Tech 
##                                      -0.0402493160 
##           factor(WordContent)2:factor(channel)Tech 
##                                       0.0195533217 
##           factor(WordContent)3:factor(channel)Tech 
##                                       0.0914589438 
##           factor(WordContent)4:factor(channel)Tech 
##                                      -0.0849238282 
##           factor(WordContent)5:factor(channel)Tech 
##                                      -0.1595821326 
##           factor(WordContent)6:factor(channel)Tech 
##                                      -0.0725825219 
##           factor(WordContent)7:factor(channel)Tech 
##                                      -0.2685540086 
##           factor(WordContent)8:factor(channel)Tech 
##                                      -0.0417582153 
##           factor(WordContent)9:factor(channel)Tech 
##                                      -0.1283152875 
##         factor(WordContent)10:factor(channel)World 
##                                      -0.5877779624 
##         factor(WordContent)11:factor(channel)World 
##                                       1.3651086489 
##          factor(WordContent)2:factor(channel)World 
##                                      -0.0027054789 
##          factor(WordContent)3:factor(channel)World 
##                                      -0.0799848362 
##          factor(WordContent)4:factor(channel)World 
##                                      -0.2258300652 
##          factor(WordContent)5:factor(channel)World 
##                                      -0.4599207830 
##          factor(WordContent)6:factor(channel)World 
##                                      -0.2996168328 
##          factor(WordContent)7:factor(channel)World 
##                                      -0.5751504127 
##          factor(WordContent)8:factor(channel)World 
##                                      -0.4304562034 
##          factor(WordContent)9:factor(channel)World 
##                                      -0.1884462999
```

The ANOVA model with the significant interaction term and the other categorical predictors was reexpressed into a multiple linear regression, which is represented above with the beta term coefficients. Having generated our first ANOVA model with the possible interaction terms in the ANOVA, we will move onto predicting the number of shares based on our ANOVA model from an article with particular traits.

Prediction

Having constructed the ANOVA model, we will carry on with the prediction of the number of shares from an article with content word count of 500 words, a positive sentiment, being published on Monday and being a Business channel. Below we determine the coefficients of the regressional equivalent of the anova model produced of the 4 factor ANOVA.

Given the above arguments in the regression equation, we retrieve the predicted number of shares based on our ANOVA model after filtering out the dummy variables which don't aren't the categories in question for prediction: $log(\widehat{shares}) = 7.180 + 0.182(WordContent.3) + 0.0713(positive_{sentiment}) - 0.044(Monday)$, where WordContent.3 = 1, positive sentiment = 1, and Monday = 1.

```
## [1] 1618.573
```

After generating the model and acknowledging many of the terms are reduced to 0 because of the dummy variable being specific to one dummy variable described in the prediction of the number of shares specific to a type of article, the predicted number of shares from an article with 500 words, a positive sentiment, being published on Monday and being a Business channel is 1618.573 article shares, or roughly 1619.

Discussion

After mining the data, analyzing the article data, and constructing a model to predict article shares, we can conclude that article shares appears to increase over the weekend in comparison to the weekdays (daypublished), on average. There's greater shares on average for positive sentiment than negative sentiment articles, articles with greater word counts tend to be correlated with greater article shares (especially after the word content category of 10), and articles in the Lifestyle channel seem to, on average, be shared more than articles from other channels such Entertainment and Business with the lowest shares which creates more distinct medians between channel groups and improves in it predicting the number of shares an article gets. In the interaction effects, the significant interactions are between the WordContent category of 10 and channel of Entertainment, WordContent category of 4 and channel of Entertainment, WordContent category of 6 and channel of Entertainment, WordContent category of 7 and channel of Entertainment, WordContent category of 8 and channel of Entertainment, WordContent category of 12 and channel of Other, WordContent category of 10 and channel of Tech, WordContent category of 5 and channel of World, WordContent category of 6 and channel of World, WordContent category 7 and channel of World. Thus, the word count in the articles affect the logarithm of the number of article shares in the channels of Entertainment, Other, Tech and World.

The response variable of the number of article shares was taken the logarithm of in the final ANOVA model because it reduced deviation of sample quantiles from normal quantiles and was the best transformation of the response variable aside from square root or other fractional power transformations to normalize the response variable. The final ANOVA model had only included the significant interaction terms because after looking through the interaction plots, the logarithm of article shares vs. WordContent per channel appeared to have interaction effects while others didn't seem to have significant interactions between its factors. The ANOVA model preceding the final ANOVA model was run with possible interactions and there was one significant interaction between the interaction between WordContent and channel. The ANOVA model predicted there to be roughly 1619 article shares given an article with 500 content, a positive sentiment, being a Business channel, and being published on Monday.

The ANOVA model converted into a multiple linear regression with an interaction term and categorical predictors had a bit of deviation from normality for lower quartiles in the data. However, we would have to acknowledge the fact that most of the response variable, article shares, were small and likely the innate extremely right-skewed nature of the article shares data lead to slightly peculiar normality in the error results. In spite of there being some very important categorical predictors used to predict the number of shares from an article such as the article sentiment, other factors could form relationships with article shares such as how much the topic is trending. This could be found in the article's topic's currency and popularity in searches online. Mashable should continuously perform data analyses on how and why their articles are shared to popularize their social media.