Exploratory Data Analysis Written Report

The BEST Fit - Philip, Olivia, Leo, Allison

2025-03-07

Introduction:

Exploratory Data Analysis:

Data Set Description:

Our project utilizes the University of California Irvine Machine Learning Repository's "Online News Popularity" data set. It includes share counts and descriptive characteristics for articles published by Mashable over two years (from 2013 to 2015). Mashable Inc. is a digital media website founded in 2005 and as of November 2015, it has over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. The data set in total, has 39644 observations, each representing an individual article. Observations include characteristics such as: Number of Words in Title/Content, Rate of Unique Words, Number of Images, Data Channel, Day Published, Rate of Positive/Negative Words, Polarity, etc. Our intention is to use the data set to predict the number of shares/virality of an article based on different variables.

Key Variables:

rate_positive_words - rate of positive words among non-neutral tokens, which captures how emotionally charged the language is.

Rate_negative_words - rate of negative words among non-neutral tokens, which captures how emotionally charged the language is.

title_sentiment_polarity - A measure of how polarizing the title is

N_tokens_content - A measure of how long the article's content is

N_tokens_title - A measure of how long the article title is

data_channel - a catagorical variable denoting article topic merged from: Data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, and data_channel_is_world.

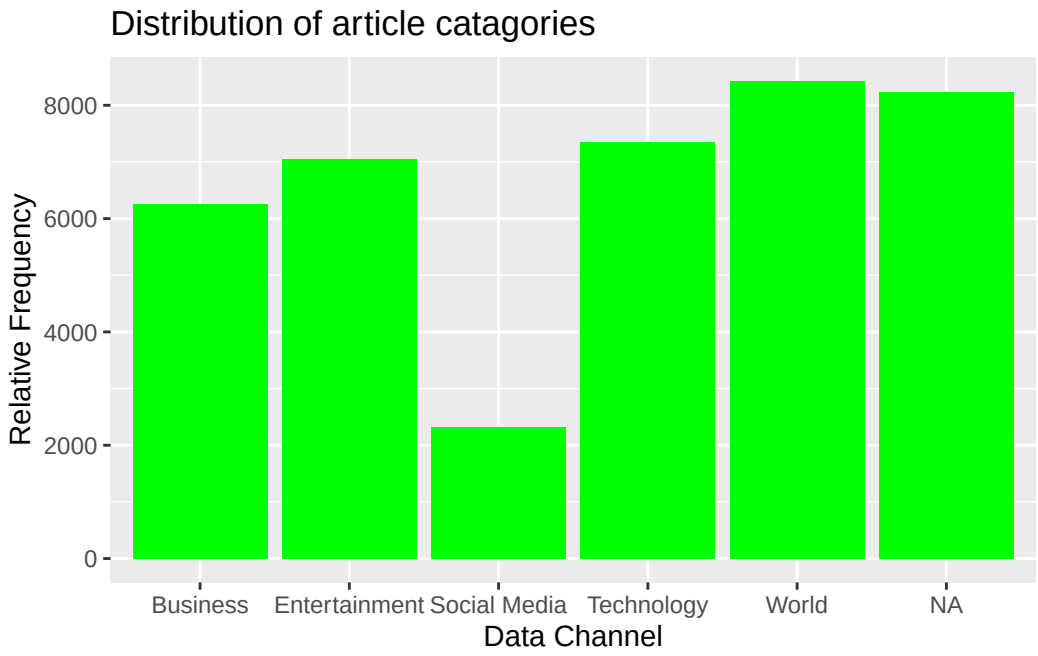
day_published- a catagorical variable indicating publication day merged from indicators: Weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday,weekday_is_saturday, weekday_is_sunday

Data Cleaning

First we have to combine the existing weekday and data_channel indicator variables into their respective categorical variables.

Table 1: Transformed Data

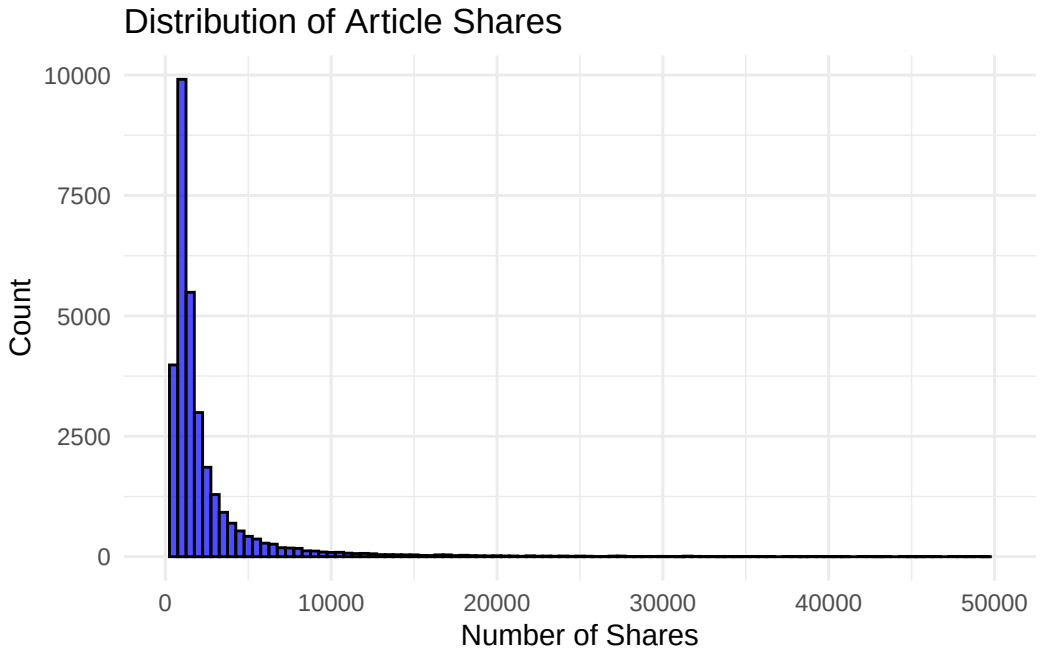
url	day_published	data_channel
http://mashable.com/2013/01/07/amazon-instant-video-browser/	Monday	Entertainment
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	Monday	Business
http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	Monday	Business
http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/	Monday	Entertainment
http://mashable.com/2013/01/07/att-u-verse-apps/	Monday	Technology



Next, we discovered that approx 8,233 articles are not tagged for a specific data channel. Due to the nature of the dataset, it's unclear if this was because the article was simply missing a tag, it was mis-tagged while being collected, or if it simply doesn't belong in any of these catagories. With the relatively large size of our dataset, we we decided to exclude entries

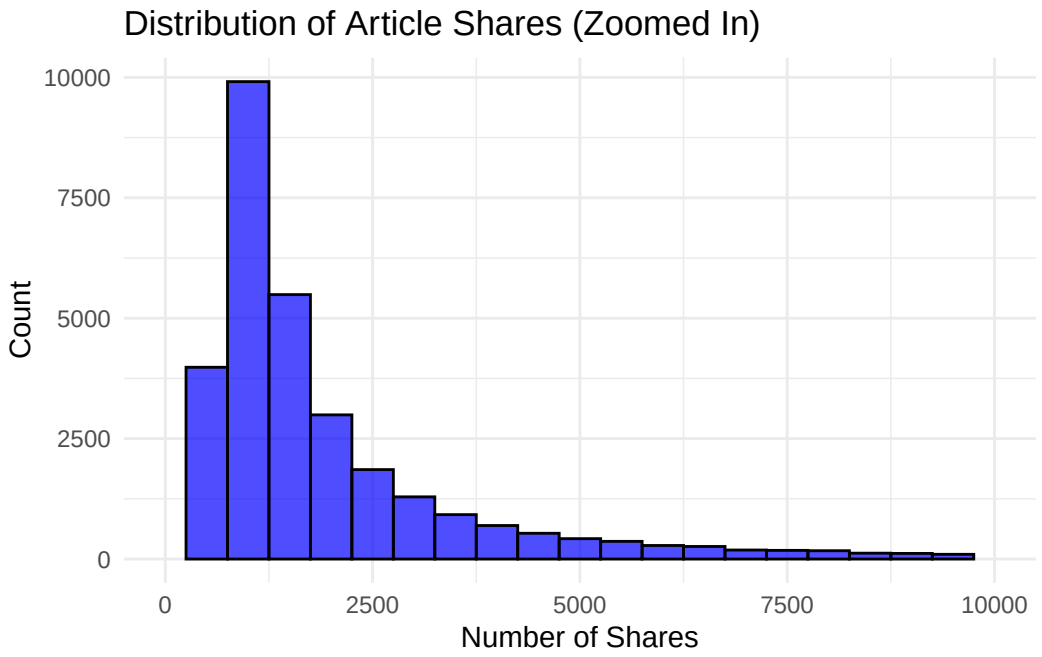
lacking a data tag NA's from our data channel analysis altogether. These articles lacking a data channel were filtered out.

Response Variable/Univariate EDA

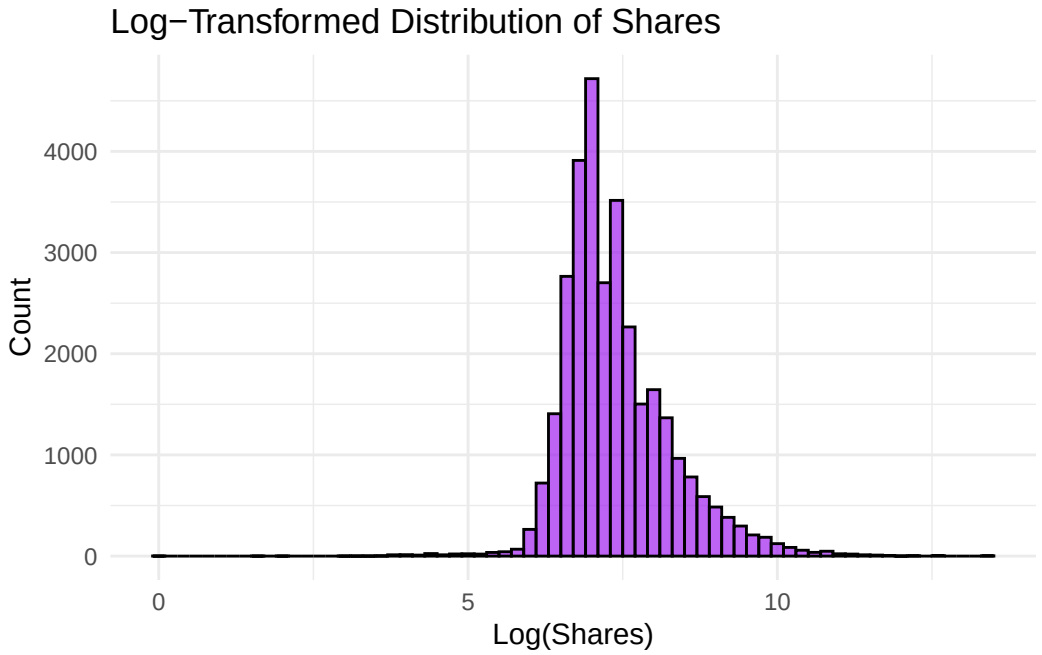



```
# A tibble: 1 x 7
  mean_shares median_shares sd_shares min_shares max_shares    q1    q3
        <dbl>         <dbl>     <dbl>      <dbl>      <dbl> <dbl> <dbl>
1       2878.          1400     9506.          1     690400   923  2500
```

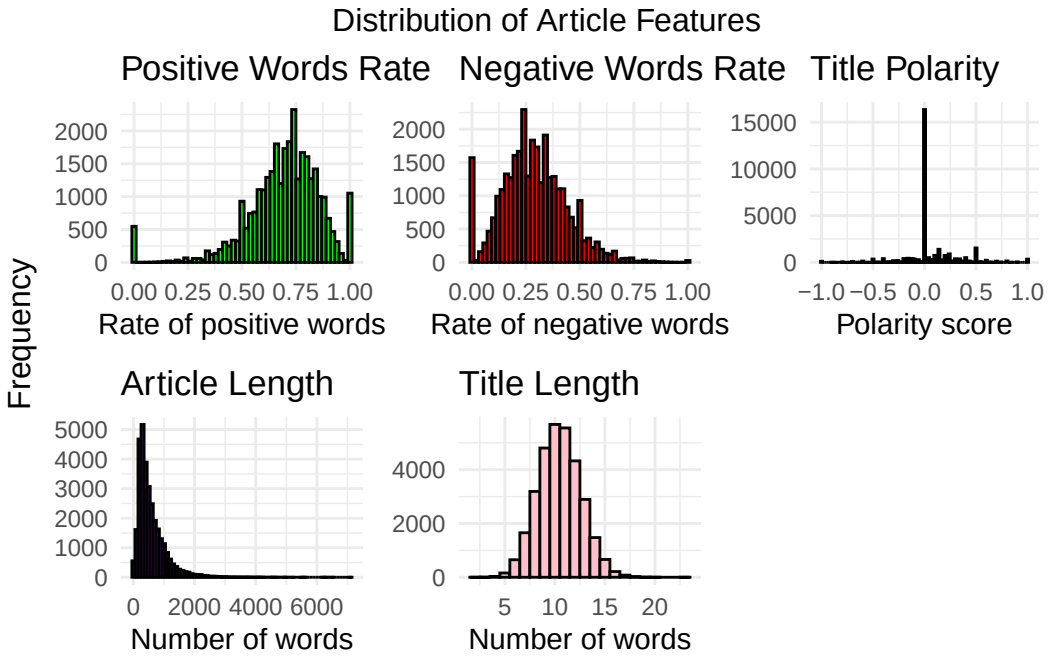
The distribution of # of shares is highly right skewed, a median of 1400 shares and a few highly shared articles. Notably, the mean of 2878 shares is far larger



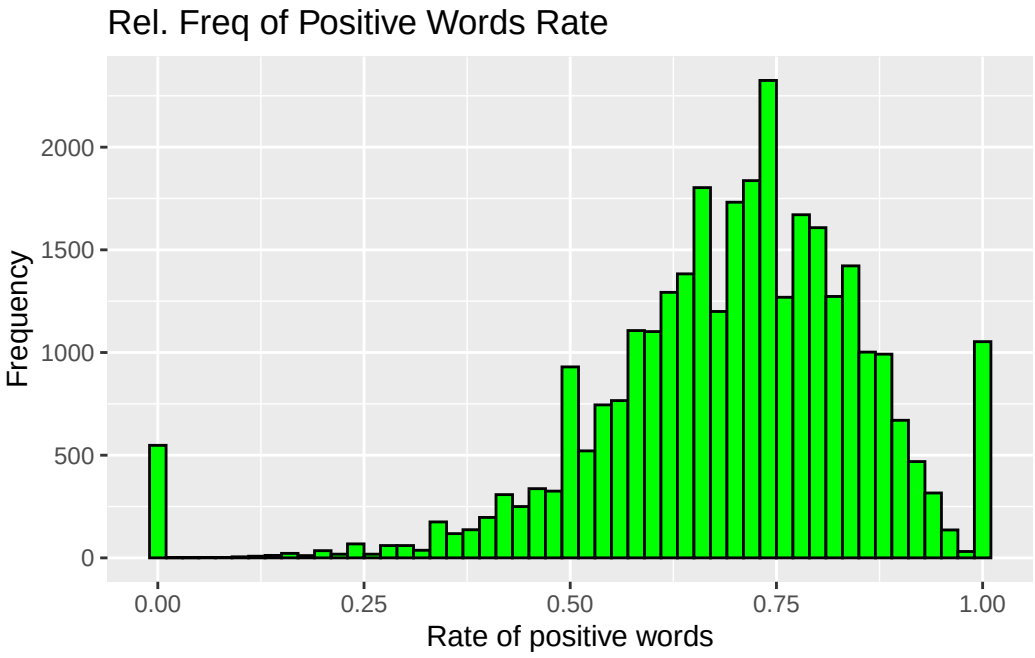
To make deal with this strong right skew, we applied a log transformation to the share variable, yielding a less skewed distribution.



Predictor Variable/Univariate EDA

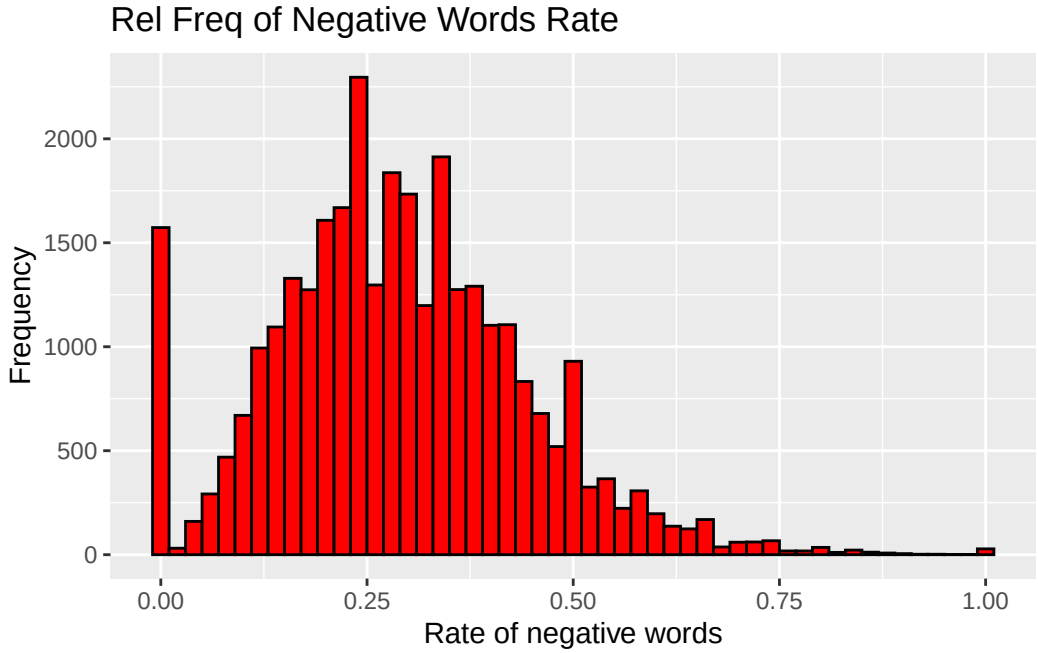


Examining the rate of positive words in an article, we see a left skew distribution, with modes at ~0, ~0.75 and ~1. The median positivity is approx 0.71 positvity rate, and the range is from 0 to 1.

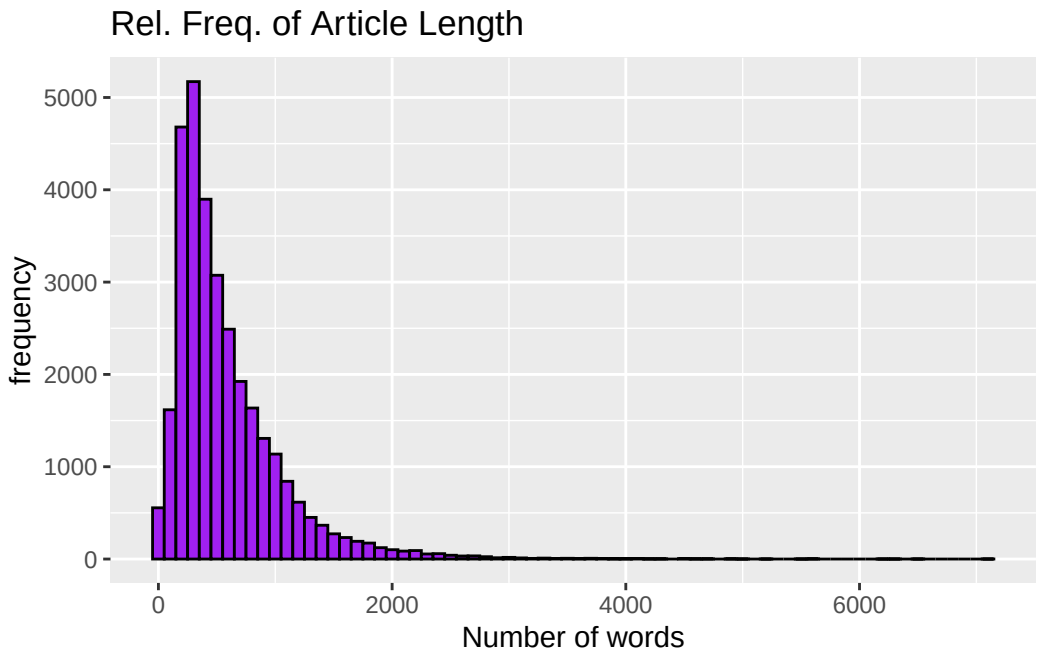


```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.694  0.714   0.172     0     1
```

The rate of negative words shows the opposite trend, with a slight right skew. Similarly, there seems to be a second mode at 0 negativity. The median is approx 0.28 negativity rate, with an approximately equal mean.

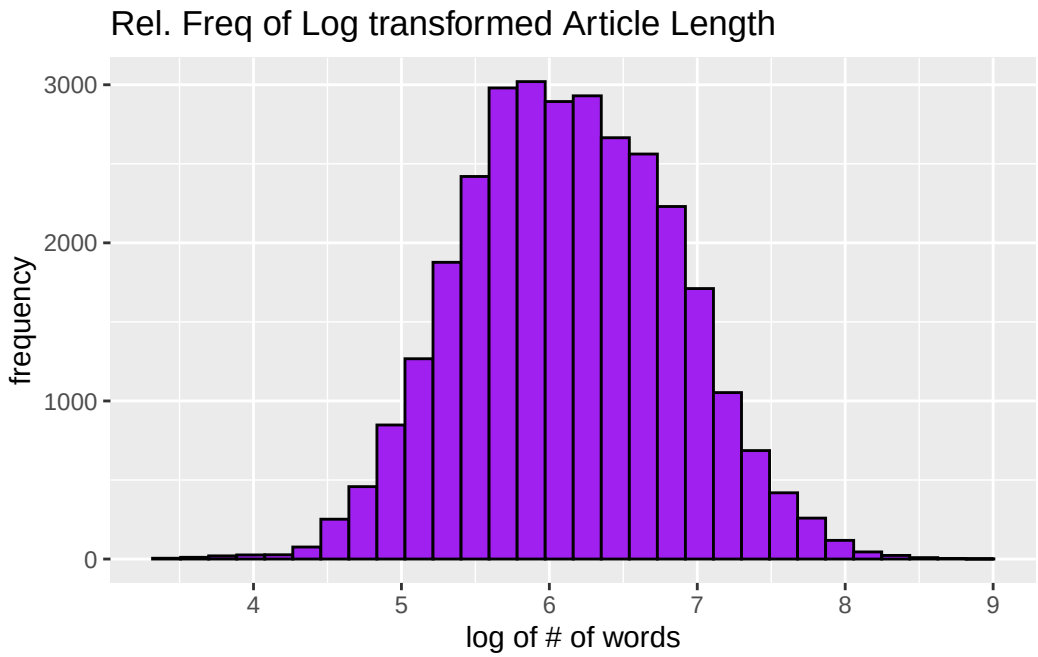



```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.290  0.280   0.152     0     1
```

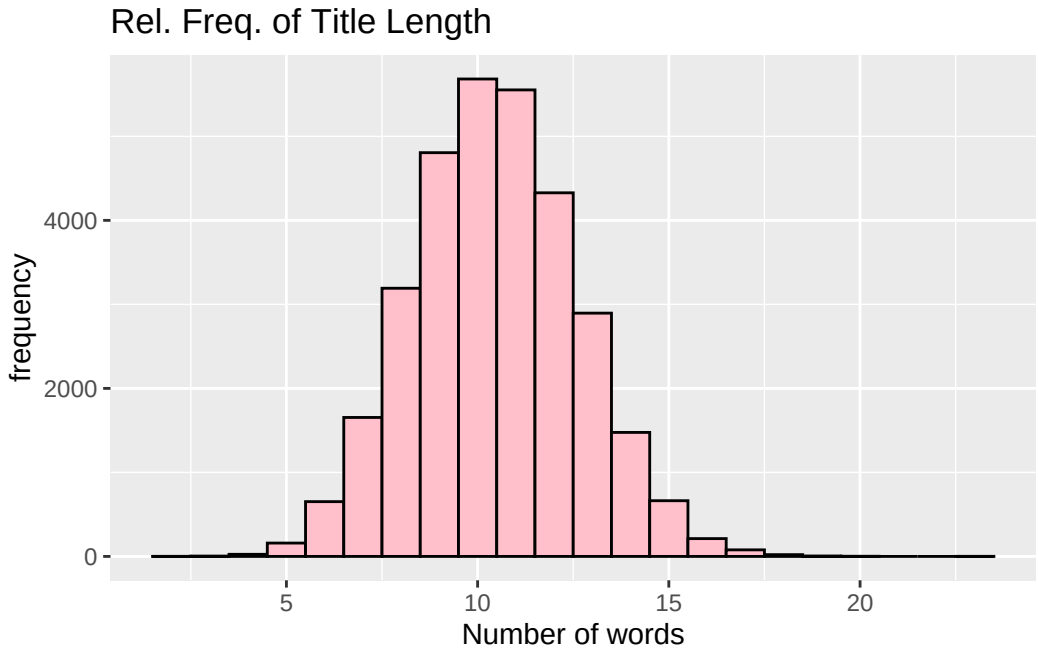



```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1  583.    444    478.     0  7081
```

For article length, the graph shows a strongly right skewed distribution, with a median length of 444 and a high standard deveation of 477 words. To remedy this, we might consider a log transfromation which yields a more even distribution.



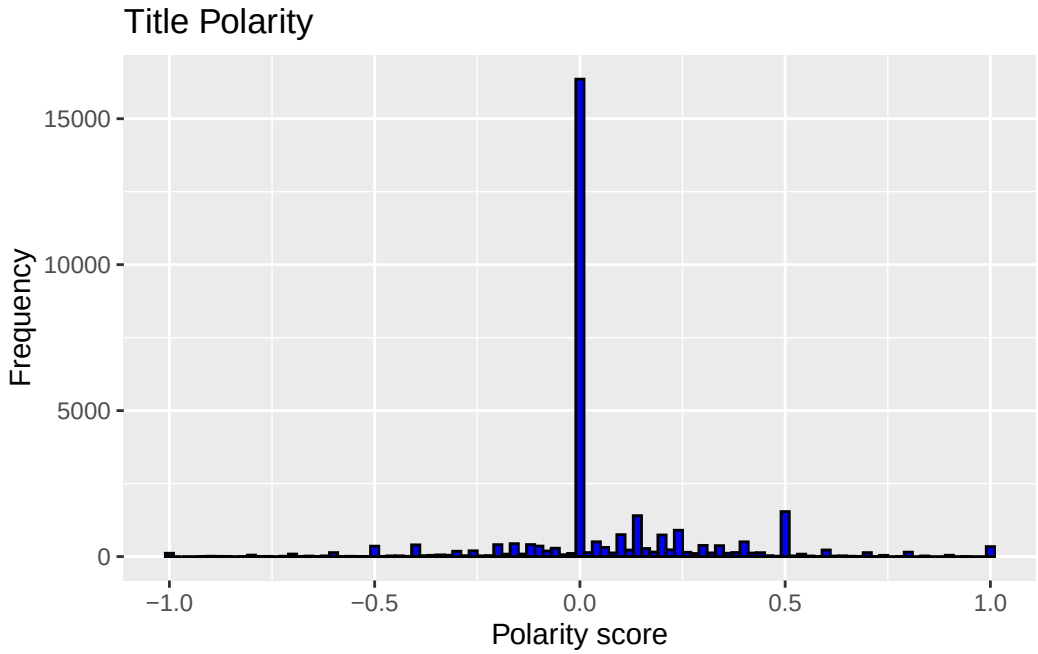
For the number of tokens in the title, we can see a highly symetric distribution centered at 10, with a standard deviation of 2.14.



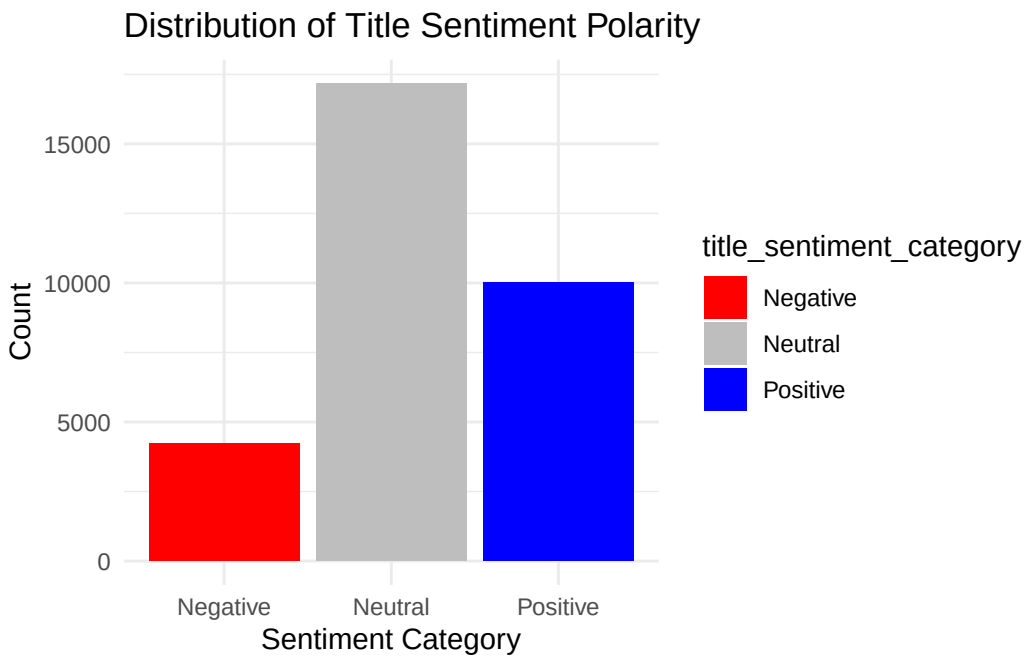

```
# A tibble: 1 x 5
```

```
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1  10.5     10    2.14     2    23
```

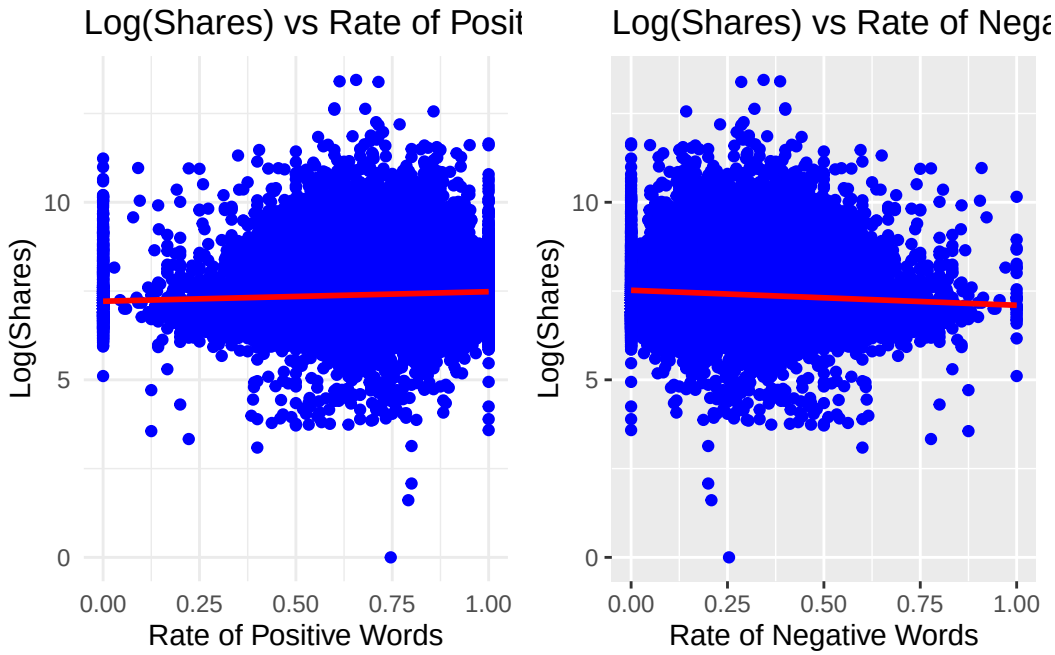
Finally, our initial EDA of title polarity found a massive frequency spike at 0 frequency, which might correspond to failed measurements or the vast majority of our articles not presenting signifigant title polarity.



To fix this issue and for ease of use, we catagorized articles into negative, neutral and positive polarity, with a threshold of 0 +- 0.05 for neutral. Most titles remain neutral, and there are more positive than negative headlines.

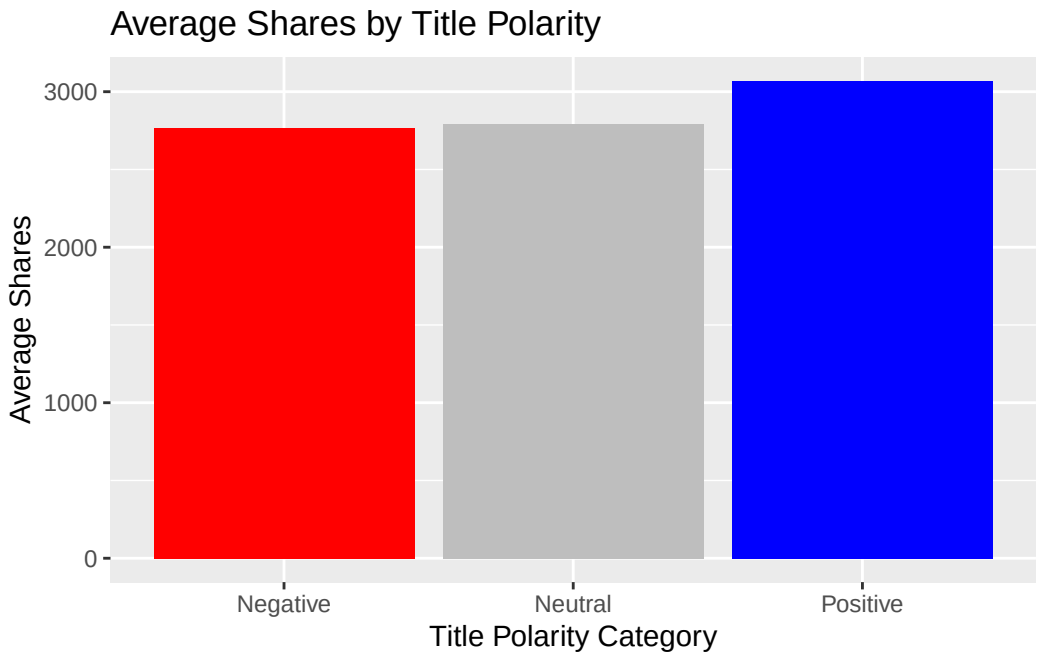


Bi-variate EDA



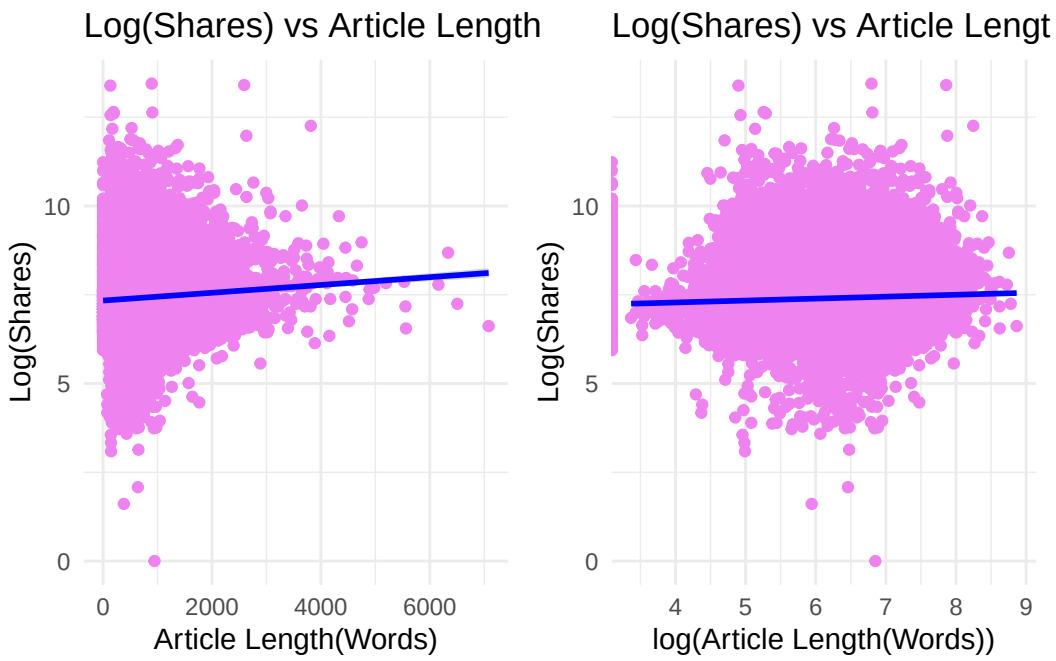
By plotting the rate of positive words and the rate of negative words against the log transformed share, we can see that neither have a particularly strong relationship with how often the article

is shared. The rate of positive words seems to have a weak positive relationship with shares, and the rate of negative words seems to have a weak negative relationship, but both have significant outliers ar 1 and 0.



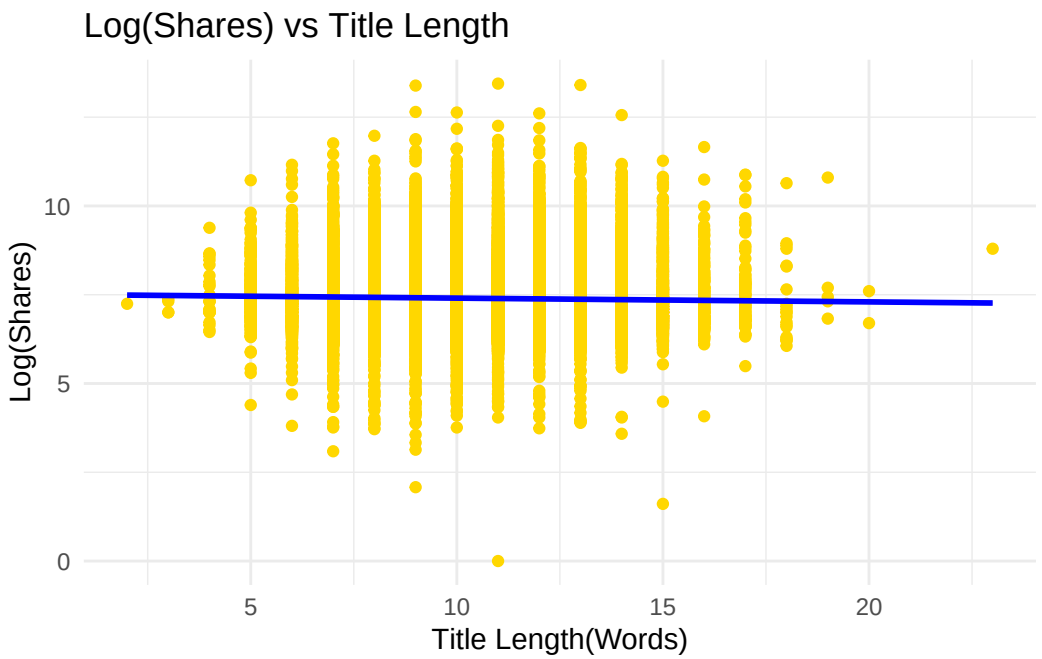
In contrast, there seems to be some relationship between title polarity and the number of shares, with positive polarity associated with greater share counts.

Next,

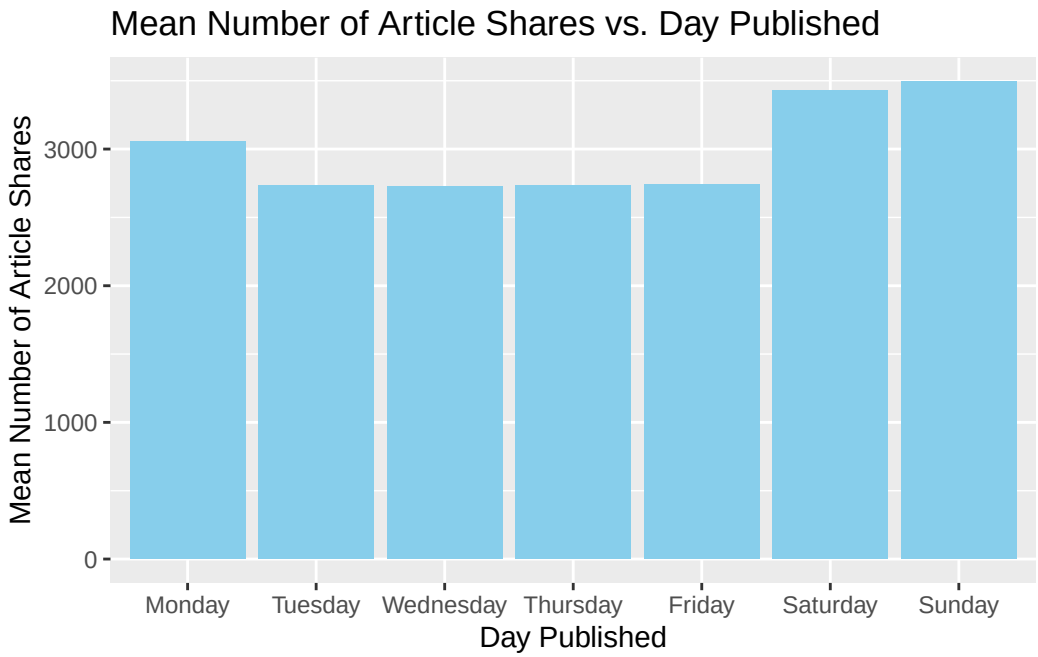


These graphs shows a weak, positive relationship betweek article length and the log transformed number of shares. Due to the skew, we can apply the log transform to the article length. This shows a more even distribution, with no clear relationship.

Similarly, there doesn't seem to be any clear relationship between title length and the number of shares in the graph below.




```
# A tibble: 6 x 2
  day_published mean_shares
  <fct>               <dbl>
1 Monday              3057.
2 Tuesday             2731.
3 Wednesday           2727.
4 Thursday            2737.
5 Friday              2741.
6 Saturday            3431.
```



term	estimate	std.error	statistic	p.value
(Intercept)	7.505	0.038	197.123	0.000
rate_negative_words	-0.408	0.045	-9.071	0.000
rate_positive_words	0.017	0.040	0.437	0.662

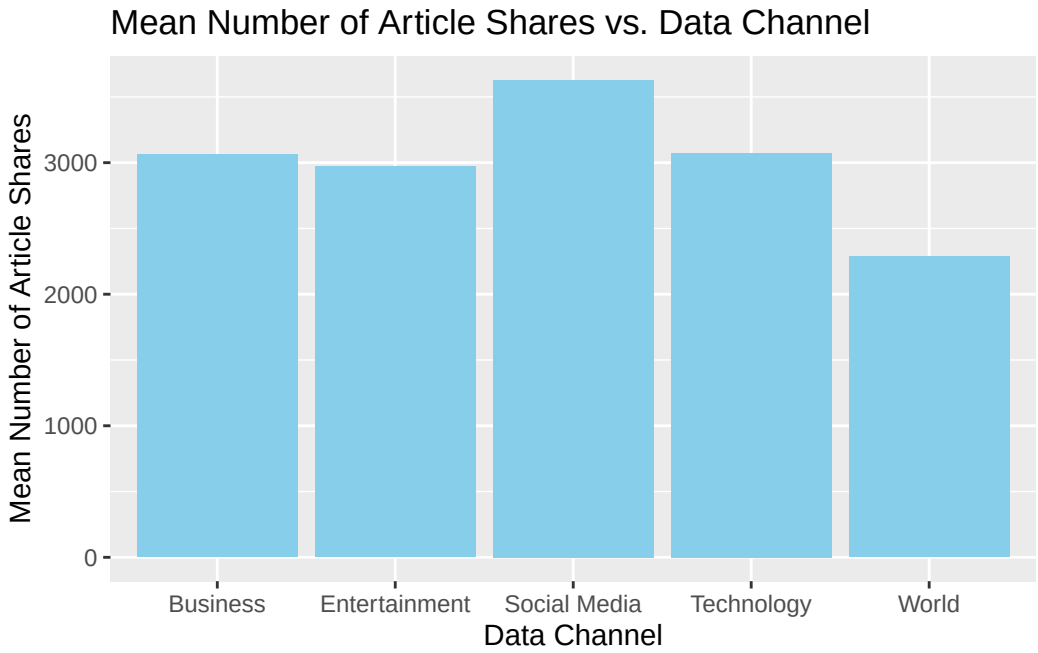
```
[1] "Variance Inflation Factors:"
```

```
rate_negative_words rate_positive_words
           1.946291            1.946291
```

This visualization and summarization suggests that the day an article is published does not have a significant impact on the virality of an article, as the mean number of article shares do not differ much between days. Therefore, this predictor may not be as important as others when it comes to predicting the virality of an article.

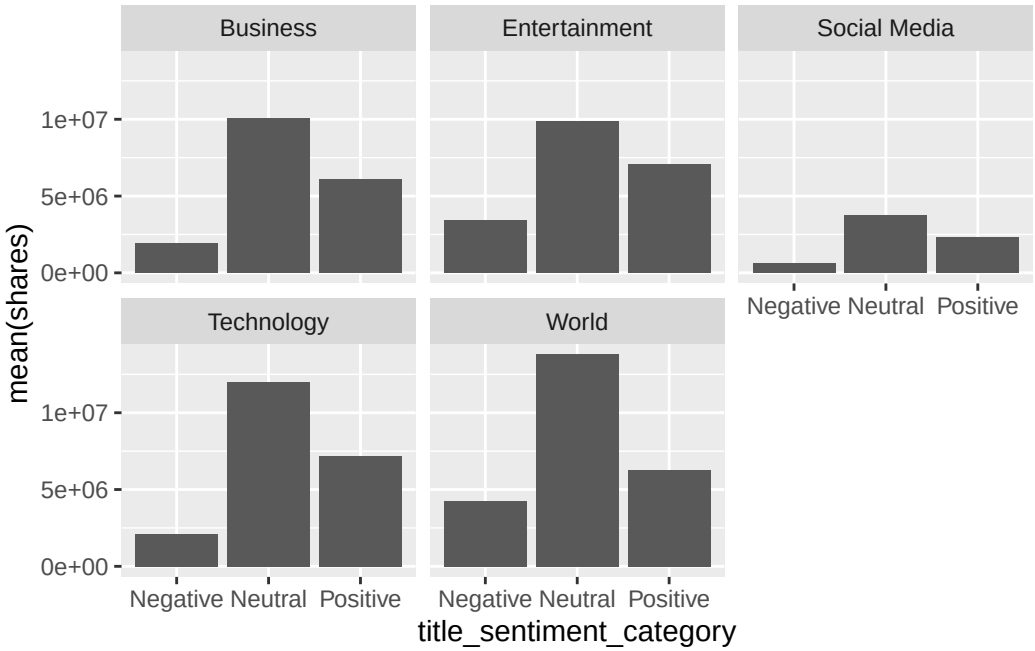
```
# A tibble: 5 x 2
  data_channel  mean_shares
  <fct>               <dbl>
1 Business            3063.
2 Entertainment       2970.
3 Social Media        3629.
```

```
4 Technology          3072.
5 World               2288.
```



From our visualizations, it appears that articles that are in the social media data channel perform the best in terms of average number of shares (~3600), while articles that are in the World data channel perform the worst (~2250). Business, Entertainment, and Technology articles all seem to receive a mean of about 3000 shares. However, from our earlier univariate analysis, we saw that the Social Media Data Channel has the lowest number of articles, and thus there is a possibility that any outliers for this data channel category would have a larger impact in skewing the mean.

Interaction Effects Exploration



From this visualization, we found that generally regardless of data channel type, articles with a mainly neutral tone experienced the greatest mean number of shares.

term	estimate	std.error	statistic	p.value
(Intercept)	7.430	0.052	141.723	0.000
n_tokens_title	-0.002	0.005	-0.417	0.677
data_channelEntertainment	-0.150	0.076	-1.978	0.048
data_channelSocial Media	0.419	0.098	4.270	0.000
data_channelTechnology	0.292	0.072	4.064	0.000
data_channelWorld	-0.415	0.071	-5.827	0.000
n_tokens_title:data_channelEntertainment	0.005	0.007	0.676	0.499
n_tokens_title:data_channelSocial Media	-0.005	0.010	-0.554	0.579
n_tokens_title:data_channelTechnology	-0.012	0.007	-1.712	0.087
n_tokens_title:data_channelWorld	0.020	0.007	2.956	0.003

From this model, we found that while the data channel type at times has a statistically significant linear relationship to log(shares), specifically for Social Media, Technology, and World, the number of words in the title does not have a linear relationship to log(shares), based on their respective p-values. For instance, the p-value for n_tokens_title is 0.677, thus suggesting that there is not a statistically significant linear relationship between n_tokens_title and log(shares). However, while the number of words in the title does not have a significant linear relationship with log(shares) directly, its interaction term specifically with when the data

channel is World, is significant (as shown by the p-value of 0.003).