Exploratory Data Analysis

Introduction

A lot of news content these days is consumed online. One way to measure the popularity of such online news articles is by counting how many likes it received or how many times it was shared.

An interesting question to consider is what factors influence the popularity of a news item. To answer this question, researchers collected data on a number of features about articles published by the digital news website Mashable for a period of two years. The dataset contains many hetereogeneous features like genre of the article, the day of the week on which it was published, the overall subjectivity of the article and it's title (scored on a range of 0 to 1 with 1 being totally subjective), the overall sentiment of the title and the article (scored on a range of -1 to 1 with -1 being totally negative), the number of images, vidoes, references for each article, the number of shares each article received and many more. The researchers' goal was to predict the number of shares an article receives in social networks based on the various features. (See the section 'Relevant papers'for more details on how the relative performance values were set.)

The full list of attribute information can be found at https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

In this project, we will look at their dataset and perform an exploratory data analysis on it with the hope of gleaning meaningful insights and uncover useful patterns.

Dataset preparation

Let us take a look at at the summary of our data set.

```
# Uncomment the line below to see the summary
# str(news)
```

Our dataset has 61 variables with 39644 entries. Let's combine some of the columns together in to categorical variables.

```
news <- data.table(news)

# Add a new column which tells what cetegory the news falls in
news[, topic := ifelse(data_channel_is_lifestyle == 1, "Lifestyle",
                ifelse(data_channel_is_entertainment == 1, "Entertainment",
                ifelse(data_channel_is_bus == 1 , "Business",
                ifelse(data_channel_is_socmed == 1, "Social media",
                ifelse(data_channel_is_tech == 1, "Technology",
                ifelse(data_channel_is_world == 1, "World", "Other"))))))]

# Add a new column which tells what day the news article was published
news[ ,day_of_publication := ifelse(weekday_is_monday == 1, "Monday",
                             ifelse(weekday_is_tuesday == 1, "Tuesday",
                             ifelse(weekday_is_wednesday == 1 , "Wednesday",
                             ifelse(weekday_is_thursday == 1, "Thursday",
                             ifelse(weekday_is_friday == 1, "Friday",
                             ifelse(weekday_is_saturday == 1, "Saturday",
                             ifelse(weekday_is_sunday == 1, "Sunday", NA)))))))]

news <- na.omit(news)
```

Univariate Plots Section

For this data set, there are many variables worth exploring. Let's dive in and first look at how the number of shares is distributed.

Number of shares

```
describe(news$shares)
```

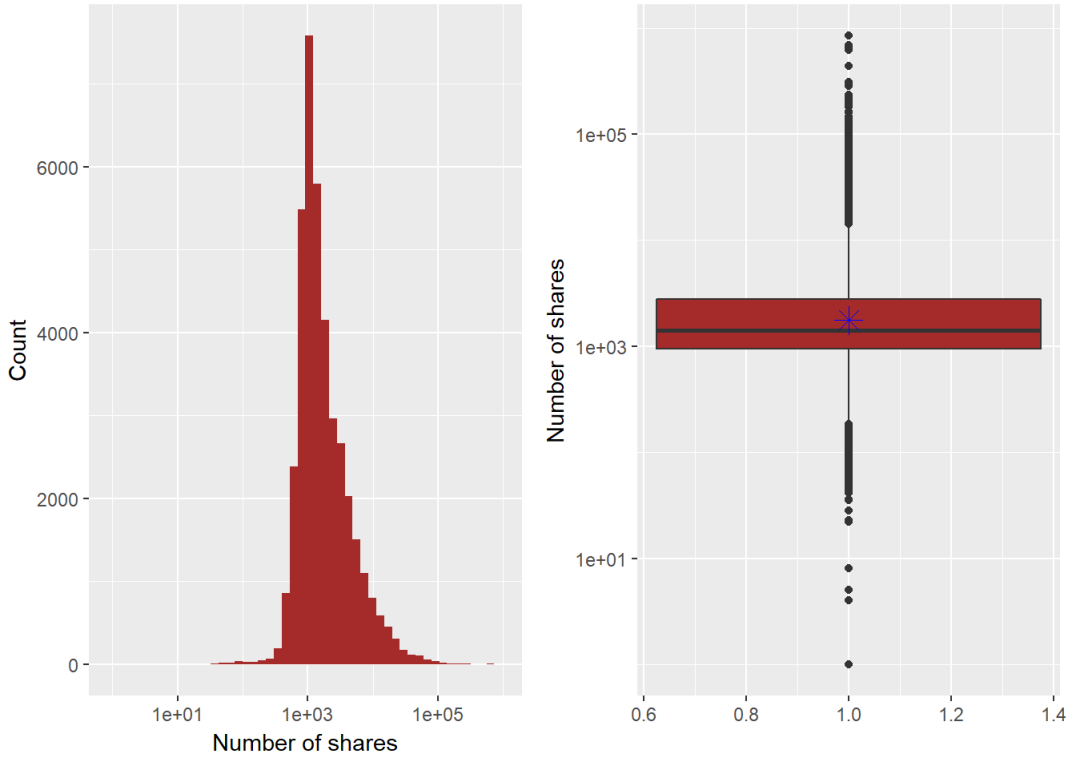
```
##    vars     n    mean       sd median trimmed    mad min    max  range
## X1    1 39644 3395.38 11626.95   1400 1888.84 889.56   1 843300 843299
##     skew kurtosis   se
## X1 33.96  1832.35 58.4
```

```
summary(news$shares)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     946    1400    3395    2800  843300
```

The mean is higher than 75th percentile implying a very skewed distribution with a lot of variability which can be confirmed by looking at the kurtosis, skewness and standard deviation. To get a better sense of the distribution, let us look at a histogram of the of shares.

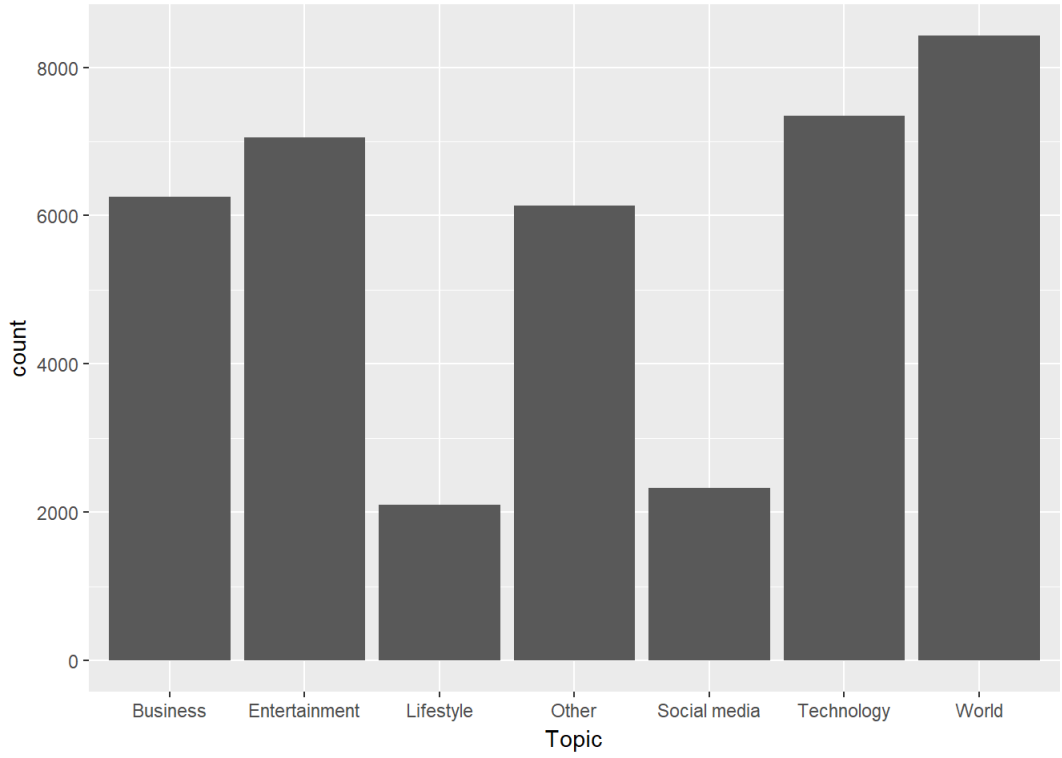
Histogram of shares

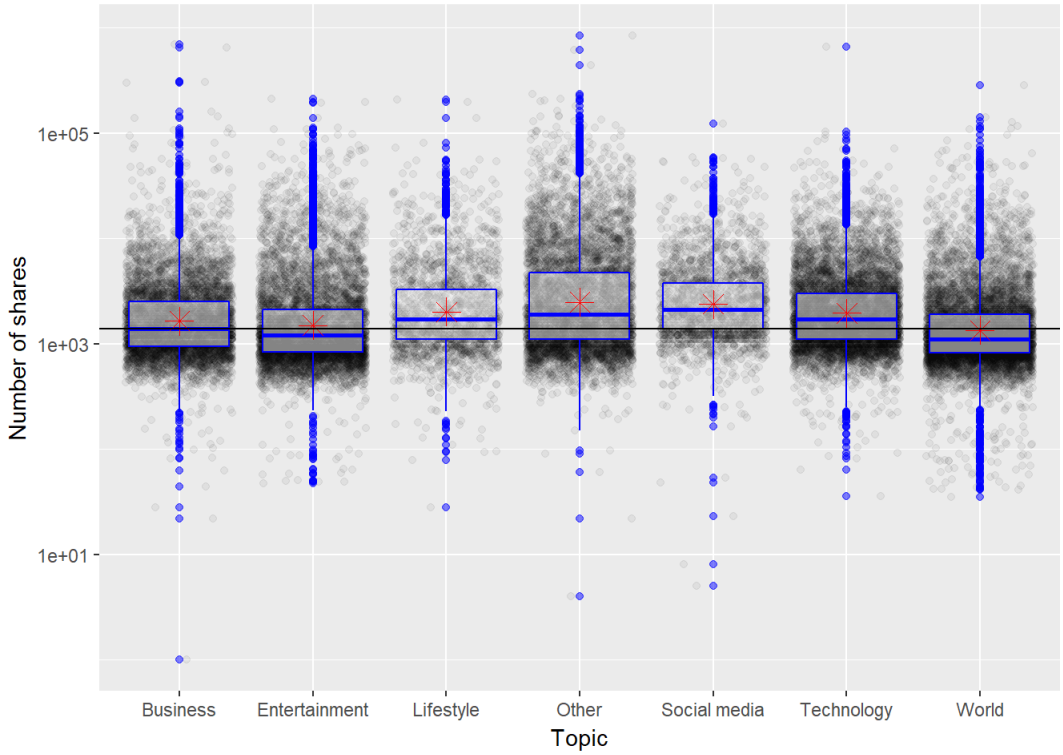


The distribution of the shares is postively skewed with very thin tails. The mean is higher than the median. There are quite a few outliers. In particular, we see many outliers with number of shares > 10000 which indicates a very high amount of popularity.

The next question we would like to explore is how does the number of shares (one metric of popularity) vary with topic? Let's look at the distribution of number of shares by topic. Before that, let's just take a look at the number of entries by topic i.e count by topic.

Number of shares by topic






```
## # A tibble: 7 × 4
##            topic median_shares avg_shares sd_shares
##            <chr>         <dbl>      <dbl>     <dbl>
## 1       Business          1400   3063.019 15046.388
## 2  Entertainment          1200   2970.487  7858.134
## 3      Lifestyle          1700   3682.123  8885.017
## 4          Other          1900   5945.190 19392.998
## 5   Social media          2100   3629.383  5524.167
## 6     Technology          1700   3072.283  9024.344
## 7          World          1100   2287.734  6089.669
```

In the second plot, the horizontal black line is the overall median number of shares. There are a couple of observations to be made. First, 'Lifestyle' and 'Social media' have far fewer entries as other topics. Second, the median number of shares for 'Social media' is highest, which is not surprising considering that Mashable started out as a blog focussing on Social media and has retained its popularity in that segment. Conversely, median number of shares for 'World' is lowest among all topics, inspite of having the most number of entries. Surprisingly, the median number of shares for 'Entertainment' is smaller than the overall median. (I was under the impression that 'Entertainment' would be the most popular. Perhaps it is the most read but not the most shared. The metric of measurement of popularity thus plays an important role. In this case, popularity is measured via number of shares.) The spread in the shares is surprisngly low for 'World' (compared to the number of articles) while it is quite high for 'Business' articles.

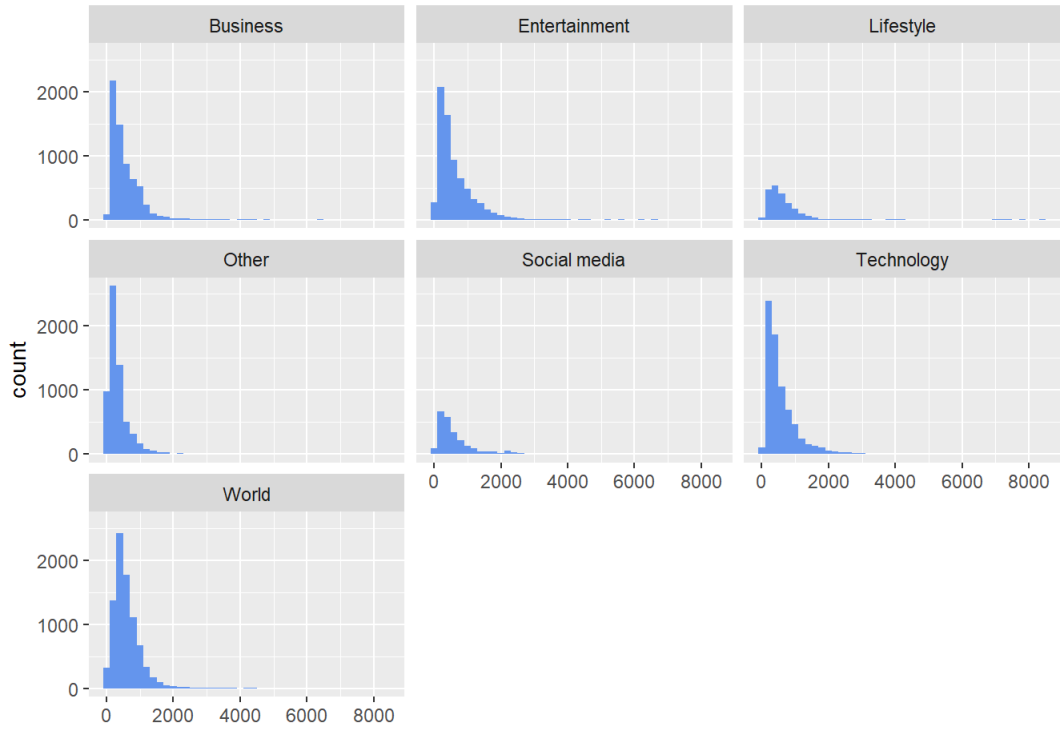
'Business' and 'World have the highest percentage of highly shared articles as the following summary shows.

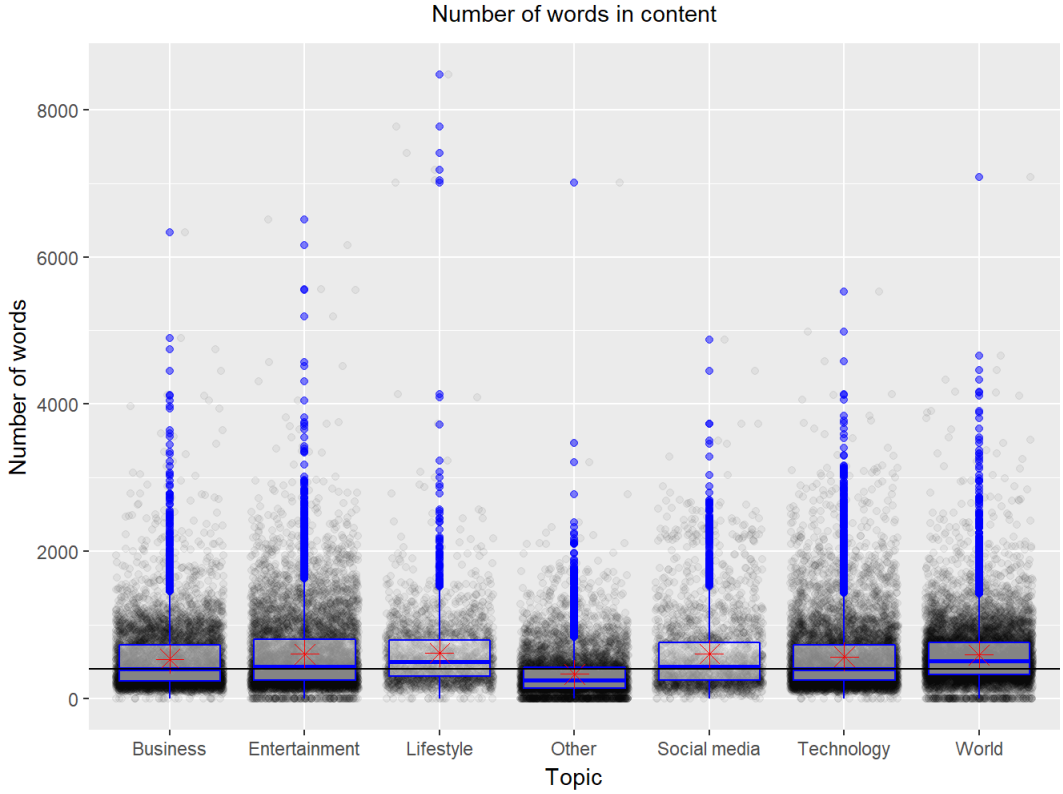
```
merge(news %>%
        group_by(topic) %>%
        summarize(total_shares = n()),
      news %>%
        group_by(topic) %>%
        filter(shares > 10000) %>%
      summarize(high_shares = n()),
      by = 'topic') %>%
  mutate(high_shares_percentage = total_shares/high_shares)
```

```
##            topic total_shares high_shares high_shares_percentage
## 1       Business         6258         193              32.424870
## 2  Entertainment         7057         389              18.141388
## 3      Lifestyle         2099         141              14.886525
## 4          Other         6134         771               7.955901
## 5   Social media         2323         130              17.869231
## 6     Technology         7346         307              23.928339
## 7          World         8427         254              33.177165
```

Word Analysis

Let us move on to word analysis and try and understand how the number of words varies by topic.





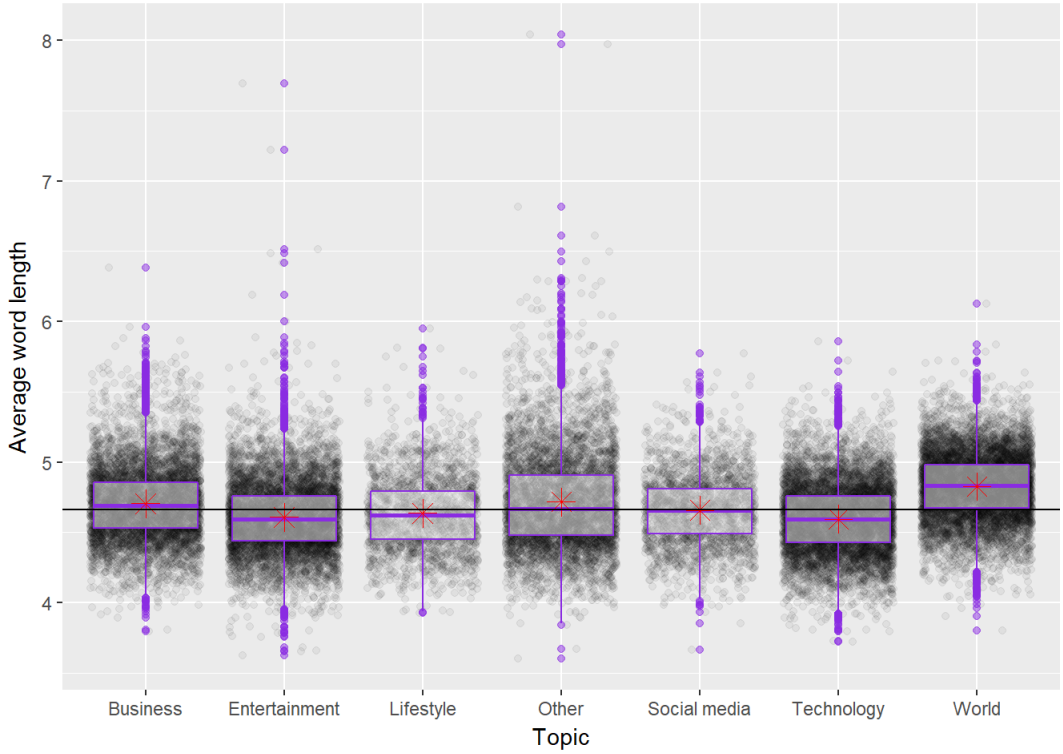
We had earlier observed that the number of entries for ‘Lifestyle’ and ‘Social’ media is the lowest. The distribution of words for different topics looks similar. We see that there are a few ‘Lifestyle’ articles which are quite wordy. Regardless of the topic, the mean is higher than the median.

The median number of words for ‘Lifestyle’ and ‘World’ are higher than the overall median. (It is probably not surpising that ‘World’ articles are, on an average, wordier, as quite a lot of them can be editorials or opinion pieces on global affairs. ‘Lifestlye’ articles being, on an average, wordier was a definite surprise. Perhaps, “Lifestyle’ articles included self-help or health related entries which could have increased the word content.

We also see a spike at 0 for many topics which implies that there are many articles with no words. This is either due to a mistake in data collection or the articles contain only images/ videos. To rule out the second possibility, here is the summary of articles by topic which have no words or videos or images. These data points most likely imply an error in data entry.

```
## # A tibble: 7 × 3
##            topic avg_shares total_count
##            <chr>      <dbl>       <int>
## 1       Business       2300           3
## 2 Entertainment       1500          13
## 3     Lifestyle       2000           5
## 4          Other       1400          48
## 5  Social media       3200           7
## 6    Technology       3100           9
## 7          World       1300          16
```

Another interesting question to consider is the variation in average word length by topic. To answer that, let us now look at the average length of words distributed by topic.



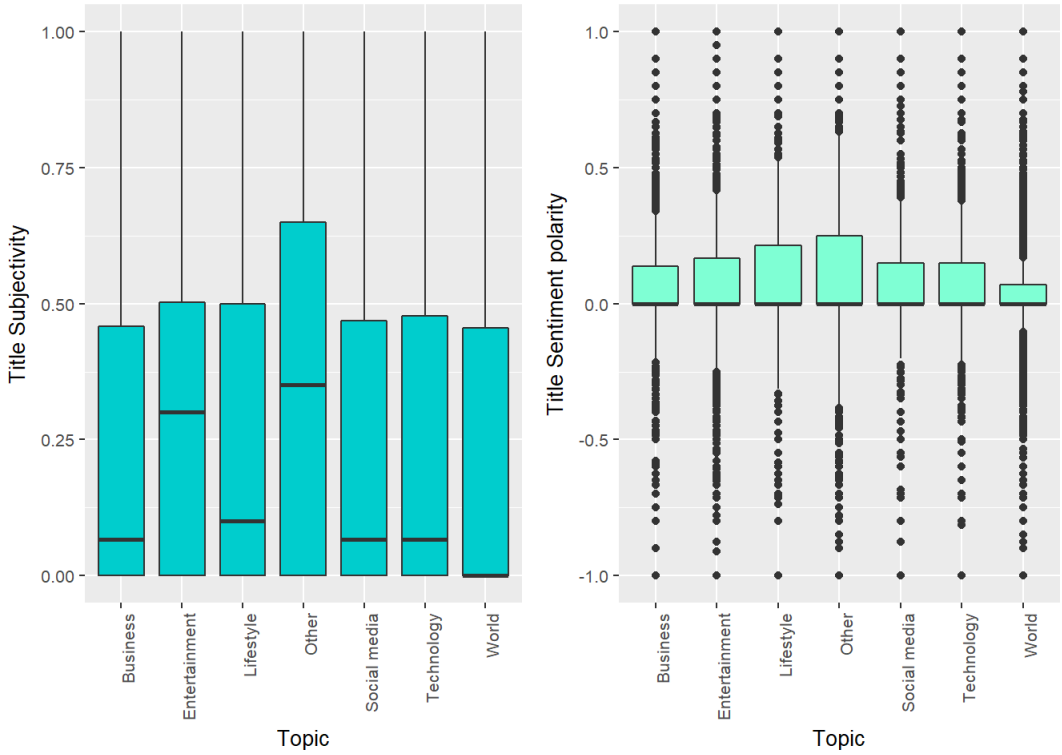
Most of the articles, across different topics have an average word length of less than 6. ‘Technology’ and ‘Entertainment’ have the least average word length while ‘World’ articles, on an average, are much more wordy with the least variation. There are a disproportionate number of ‘World’ articles with an average word length of more than 5, as the following summary shows (perhaps giving a ‘scholarly’ feel to such articles.)

```
## # A tibble: 7 × 2
##            topic Total
##            <chr> <int>
## 1       Business   763
## 2 Entertainment   444
## 3     Lifestyle   197
## 4          Other   987
## 5  Social media   201
## 6    Technology   397
## 7          World  1806
```

Let’s now move on to analyze how the title tone and subjectivity varies by topic.

Title analysis

Title of an article is the first thing a reader sees. Let’s explore the subjectivity and sentiment polarity of the titles distributed by topic.



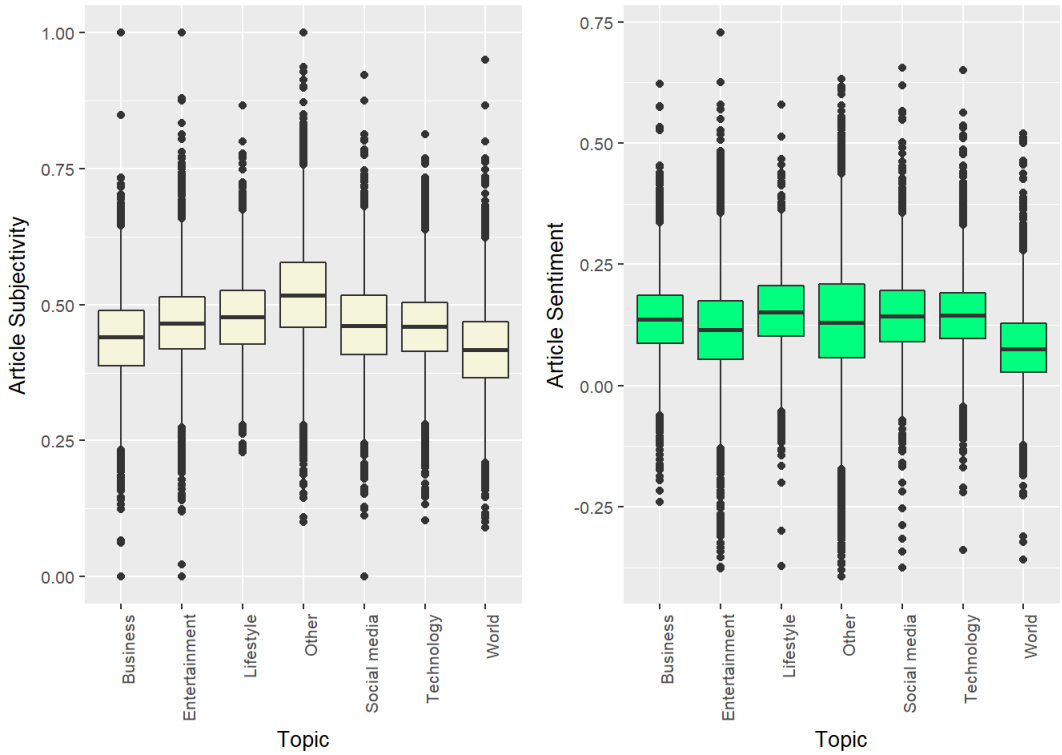
There are a few things that stand out. There are no outliers in the Title subjectivity boxplot. The median and spread for title subjectivity for 'Entertainment'is quite high while that for 'World' is lowest. One can perhaps conjecture that 'World' news article titles are usually more objective and are less likely to be 'click-baity'. (We can test this hypothesis by further looking at the article subjectivity.)

The median title sentiment polarity for every topic is zero with lots of outliers. Moreover, the distribution for each topic looks symmetric. We also observe that the spread (IQR) for 'World' is the narrowest.

Article analysis

Let's now understand how the article tone, subjectivity and sentiment polarity varies by topic.

Subjectivity and sentiment polarity

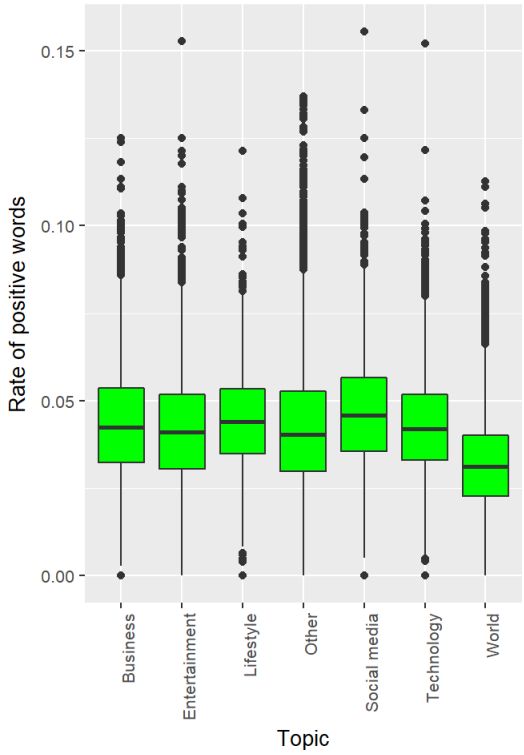


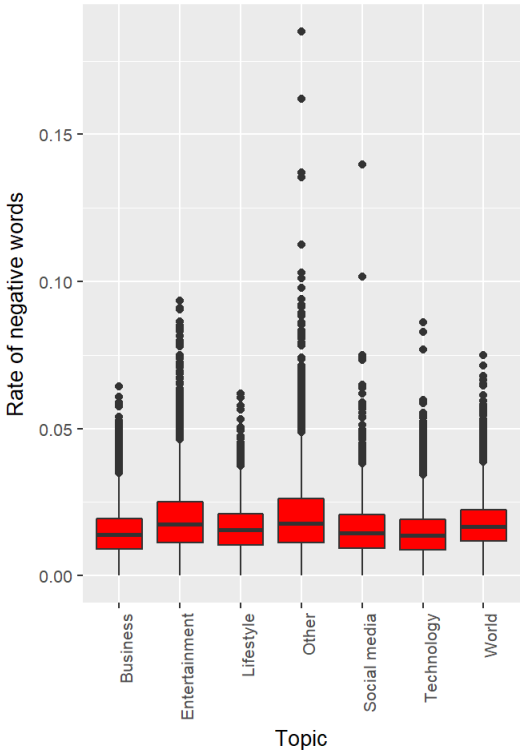
On an average, articles across topics are quite objective with 'World' articles being most objective, while articles in the 'Other' section (Sports, Editorials etc) are most subjective.

Article sentiment polarity is again the least for 'World' giving more support to our conjecture that 'World' articles are more measured in their tone and outlook. An interesting observation is that the lowest article sentiment polarity across any topic is more than -0.4 (compared to lowest title sentiment polarity of -1). The overall outlook of the published articles on Mashable is never too negative!

Let us now study the article tone by looking at the rate of positive and negative words in the content.

Tone analysis

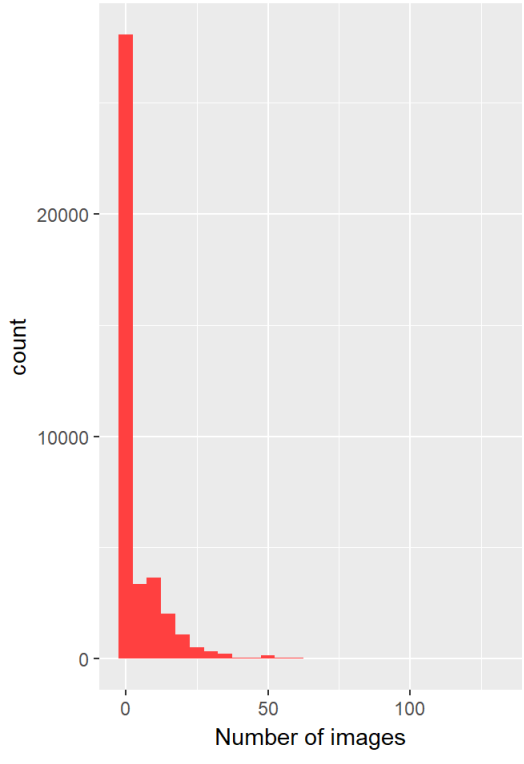


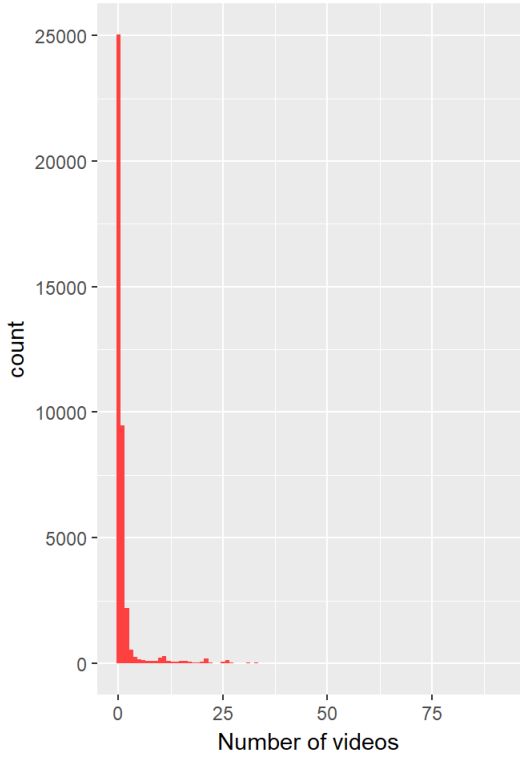


There are two immediate observations that can be made. First, the median and the spread of rate of negative words is much smaller than for rate of positive words across all topics. This further supports our previous observation that, overall, the content of articles on Mashable has a 'positive' spin to it. Moreover, the 'World' articles have the lowest postive word rate median and the difference between the postive and negative word rate median for 'World' articles is the the smallest. This is more evidence that 'World' articles, on an average, tend to be more nuanced.

Number of videos and images.

Another aspect worth looking in to is the the distribution of number of images and videos in the news articles which we now proceed to do.





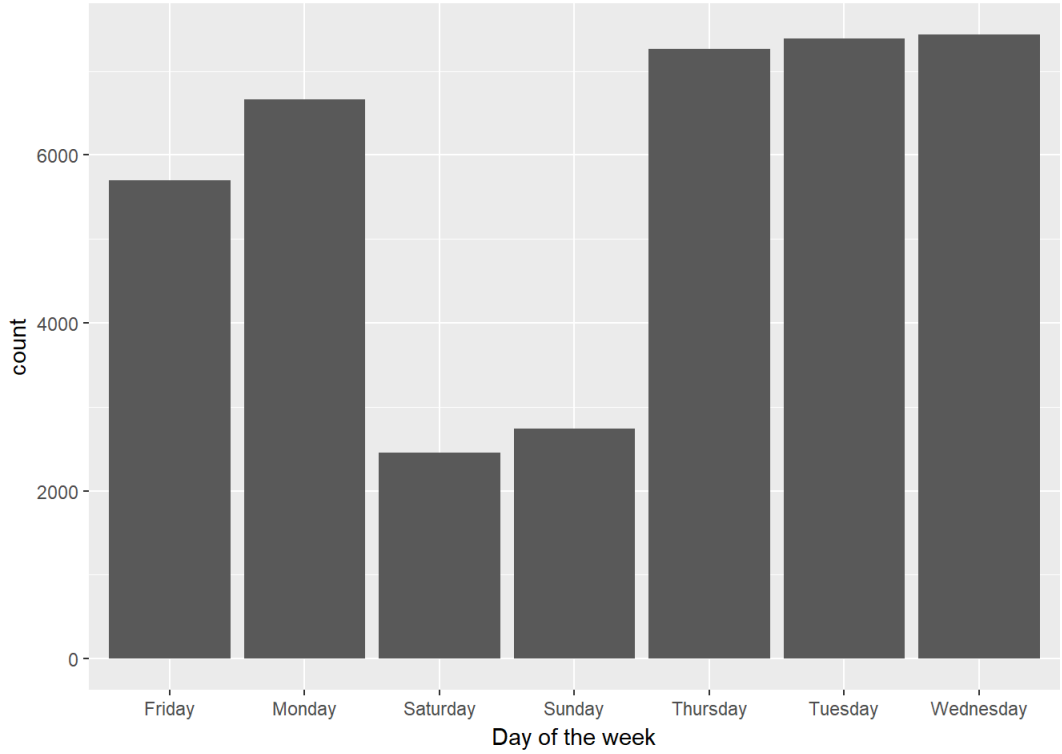
We observe a huge spike near 0 for both the number of images and vidoes implying that most articles have atmost a one or two images/ videos. Moreover, the distribution of number of images is more spread out than the distribution of number of videos.

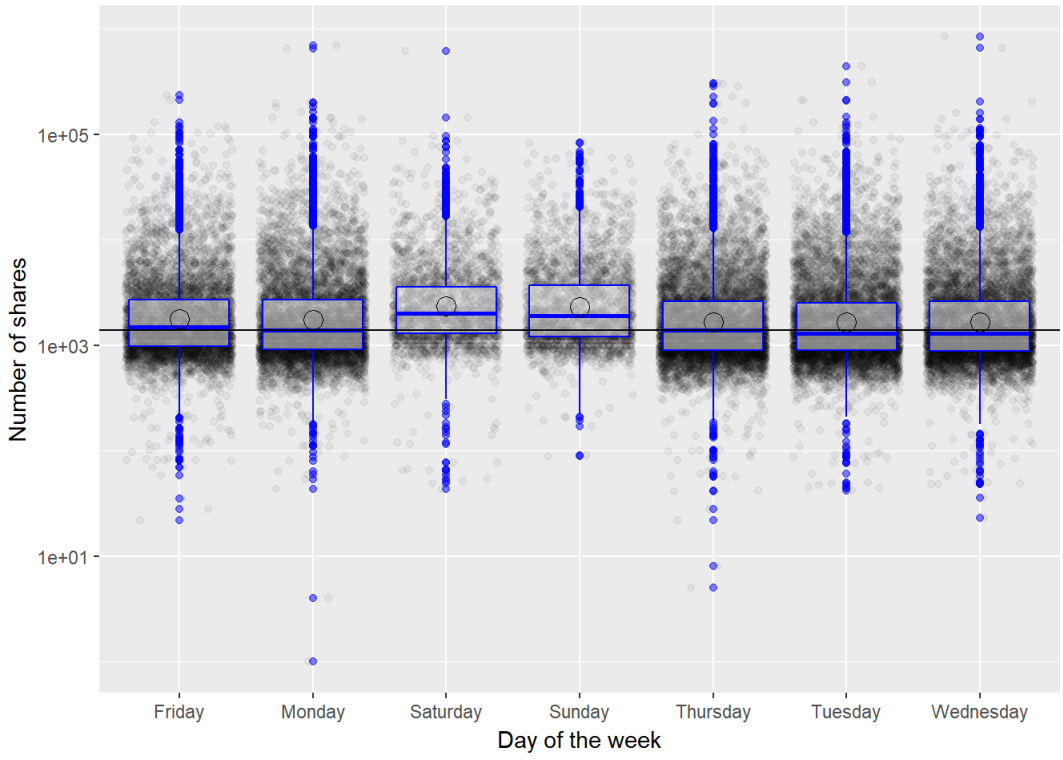
Normally, 'Entertainment' sections have more videos, on an average, than other genre of articles. Let's verify if this suspicion is true for articles published on Mashable.

```
## # A tibble: 7 × 4
##           topic avg_images avg_videos total
##           <chr>      <dbl>      <dbl> <int>
## 1      Business          1          0  6258
## 2 Entertainment          1          1  7057
## 3     Lifestyle          1          0  2099
## 4         Other          3          1  6134
## 5  Social media          1          0  2323
## 6    Technology          1          0  7346
## 7         World          1          0  8427
```

It indeed is.

So far we split the data by topic. For the final plot in this section, let us just look at number of shares according to the day of publication.





The horizontal line is the overall median of shares. It becomes clear that far fewer articles are published on weekends but quite surpisingly the median number of shares is much higher. This can perhaps be attributed to readers having more free time to read (and hence share) news on the weekends.

It is now time to summarize all that we have learned so far from the univariate plots.

Univariate Analysis

We first looked at the distribution of number of shares and found out it is highly positively skewed with thin tails with median number of shares at 1400. Next, we studied the variance in number of words and average word length for different topics. We found that are a few entries which are most likely mistakes. We also observed that 'World' articles on an average are much more wordy and have a 'scholarly' feel.

Next, we looked at the outlook and tone of the titles of articles distributed by topic. Based on the median and spread, we hypothesized that 'World' news article titles are usually more objective and are less likely to be click-baits. We then proceeded to look at the overall tone, sentimentality and outlook of the articles distributed by topic. We found that 'World' news articles are more objective in their approach, more nuanced in their tone and more measured in their sentimentality.

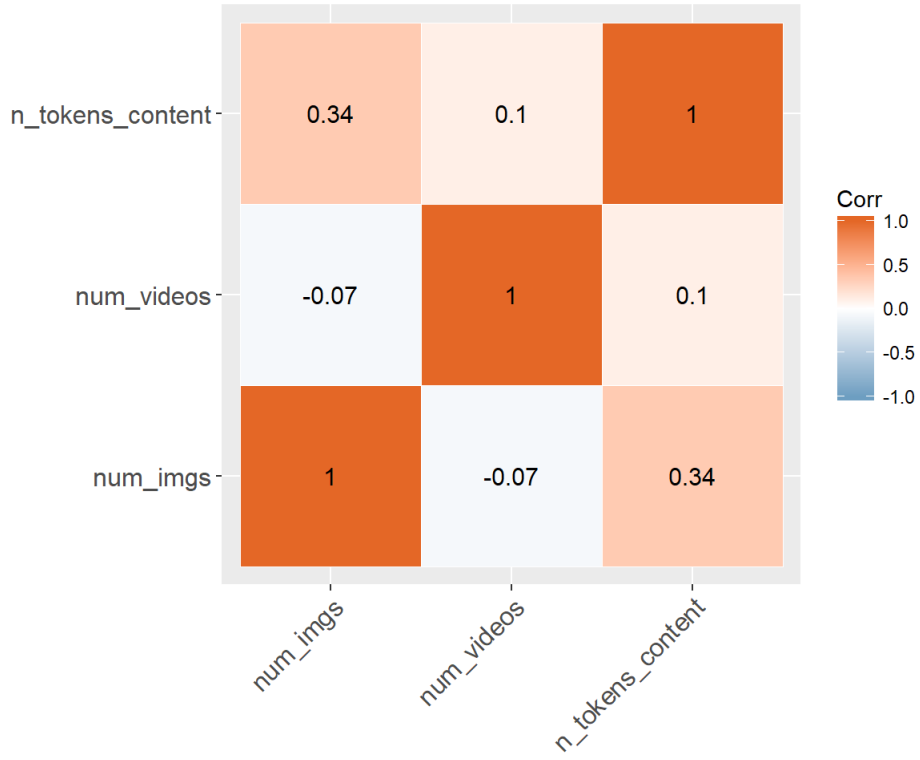
Next, we looked at the distribution of number of videos and images and found out that most of the articles have fewer than 2 images or vidoes. The distribution of number of videos is tighter with very few articles having more than 5 vidoes (which is expected as Mashable is primarily a digital media website and not a video log website.)

Finally, we looked at the median number of shares by day of publication and found out that far fewer articles are published on weekends but the popularity of articles published on weekends is much higher.

Bivariate Plots Section

Based on what we saw in the univariate plots, let's explore relationships between different variables. In the previous section, we mainly concentrated on the behaviour of different variables across various topics. In this section, let's choose one of the variables studied previosly i.e shares and see how the number of shares is affected by different variables.

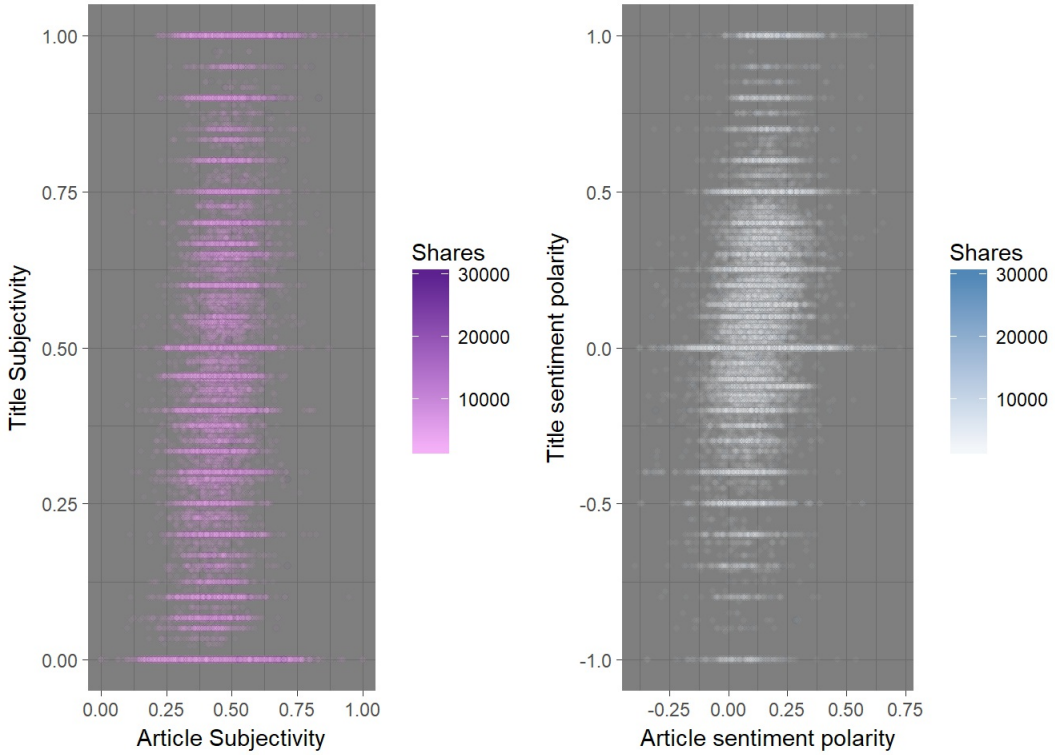
Before we go on to plots, here is a figure showing the correlation beween some of the variables we will be studying in this section.



There does not seem to be any significant (linear) relation between the number of videos or images. Number of words in an article and number of words in articles are somewhat positive correlated. The correlation between other groups of similar variables is shown in the 'Multivariate Plots' section.

Let us plot a scatterplot of title subjectivity/sentimentality versus article subjectivity/sentimentality and see how the number of shares is influenced by both. (Because there is a lot of skewness in the number of shares, we will only look at the bottom 99 percentile of the shares for this plot.)

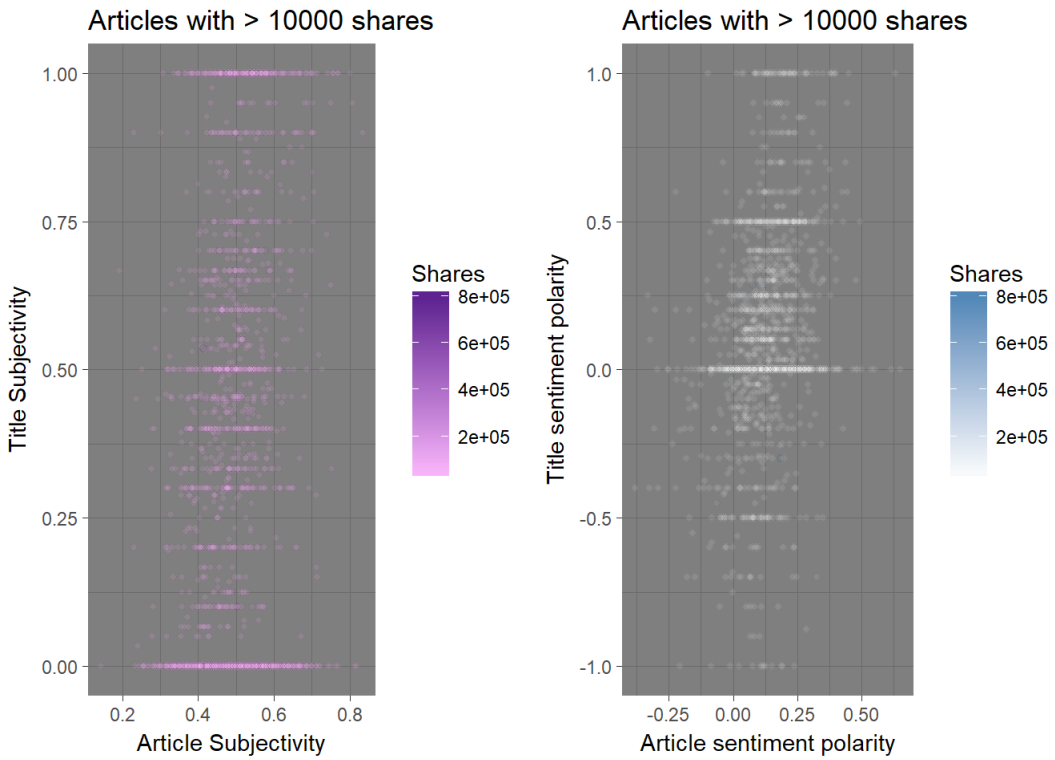
Effect of tone and content of article and title on number of shares



We notice frequent horizontal bands in the left plot implying that there are clusters of values for title subjectivity at which many entries are made. Moreover, we see that for title subjectivity = 0 or 1 (corresponding to most objective and least objective respectively), the number of shares is quite high.

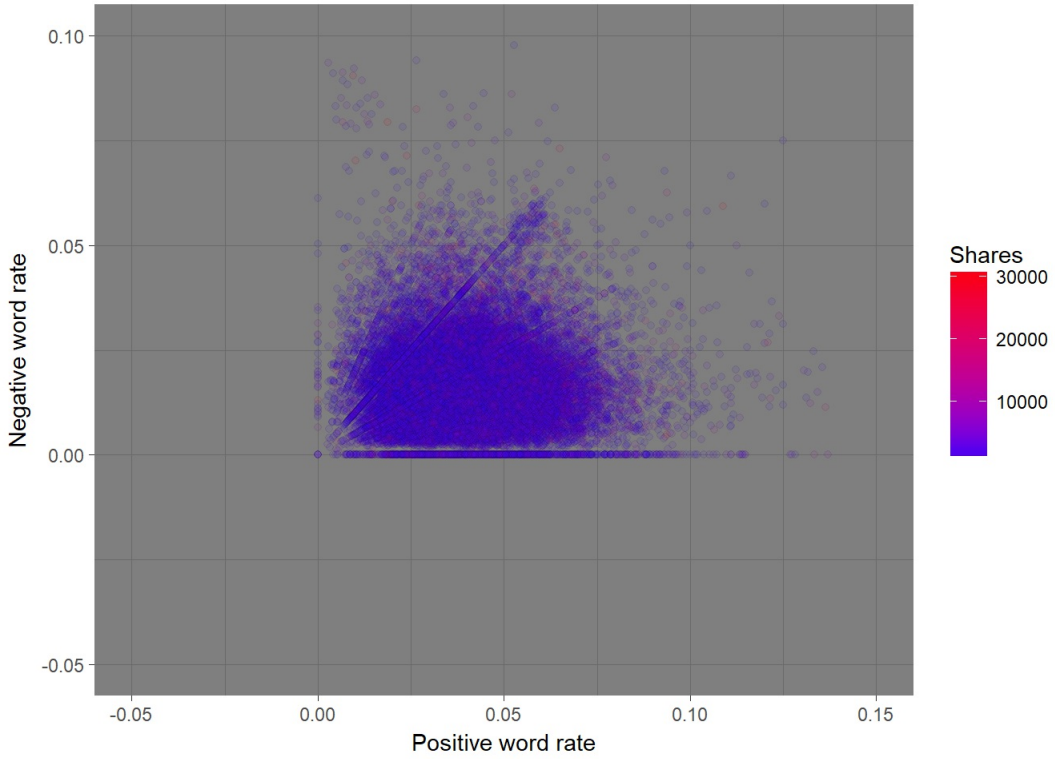
A similar pattern is observed in the right plot too. We see clusters of values for title sentiment polarity at which many entries are made. The number of shares is high for title sentiment polarity around 0.

One can perhaps conjecture that for articles which have high shares, the titles are either very subjective or very objective (1 or 0) and title sentiment polarity is around 0 (neutral). Let's explore this further by looking at a scatterplot for only those articles which have a very high number of shares, say 10000.

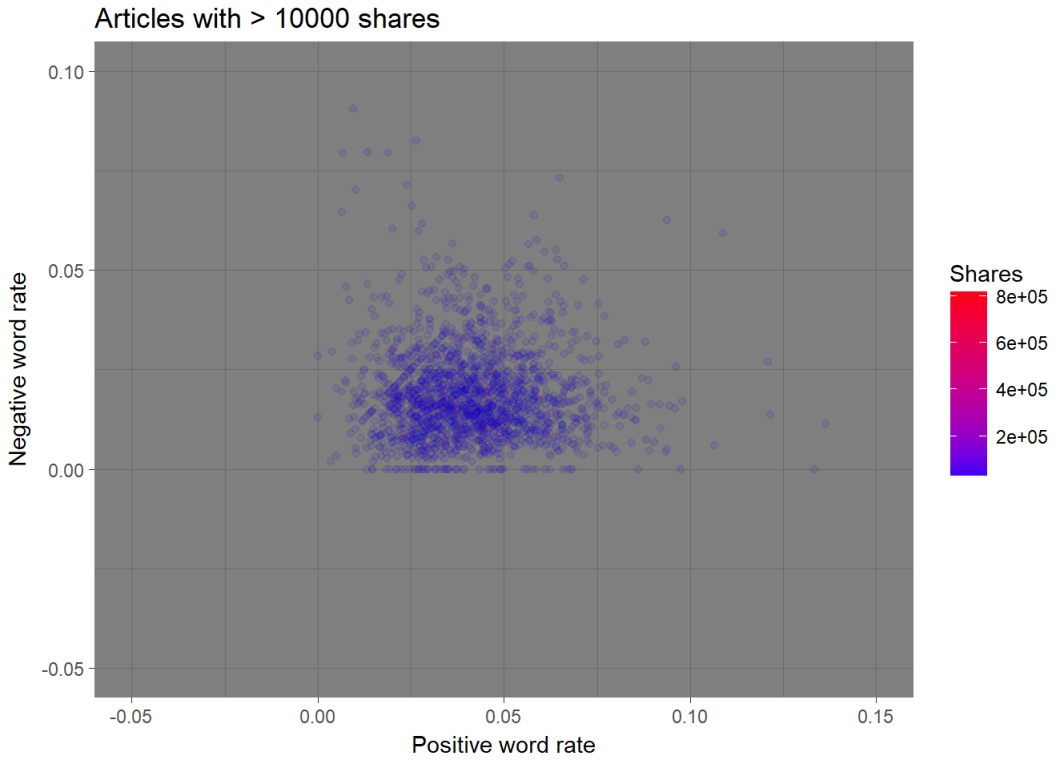


As we correctly suspected, articles with higher shares tend to have either very subjective or very objective titles and tend to have neutral sentiment polarity.

Let's now look at the influeunce of the outlook of the article on the number of shares. Again, we will only look at the bottom 99 percentile of the shares for this plot.

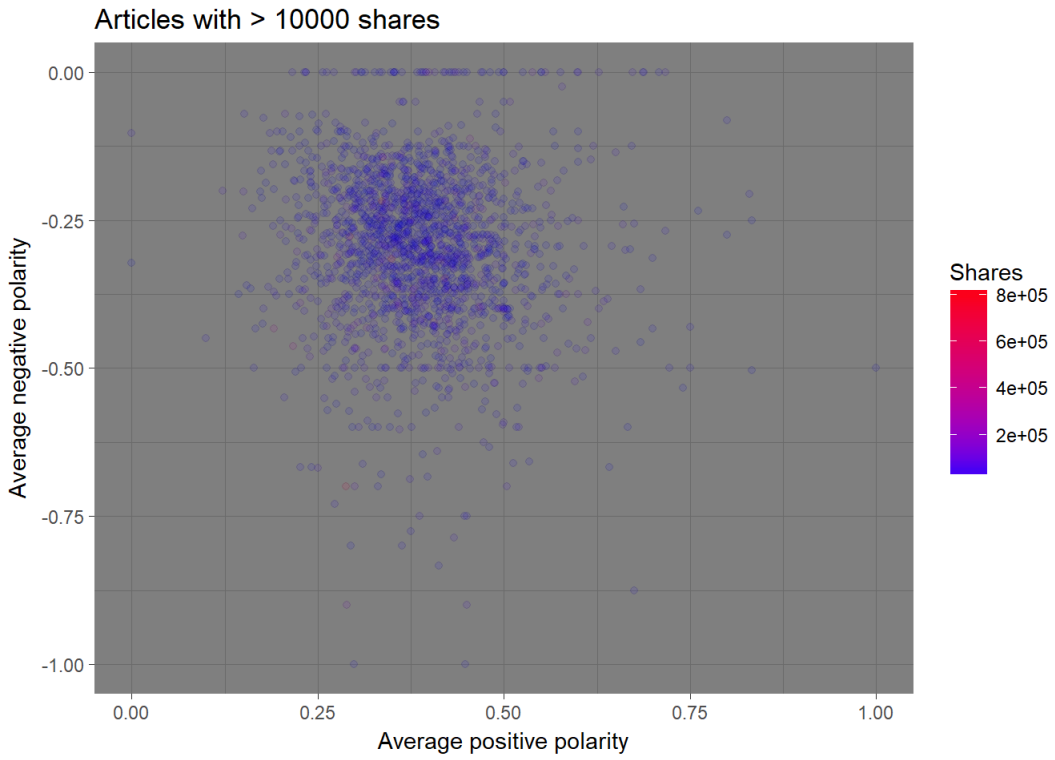


We observe a dense concentration of entries in a particular area of the plot. We also see faint linear trends (with different slopes). Not much can be discerned about the effect of the positive and negative word rate on the number of shares from this plot. Let's filter our data and only consider those entries which have more than 10000 shares.



Most of the entries with more than 10000 shares have rate of postive words <= 0.075 and rate of negative words <= 0.05 which was true for other lower shared entries too, as seen in the previous plot. From this plot, we can say that for articles with very high number of shares, the global positive word rate tends to be higher than the global negative word rate. (Notice the concentration of points below the line y=x.)

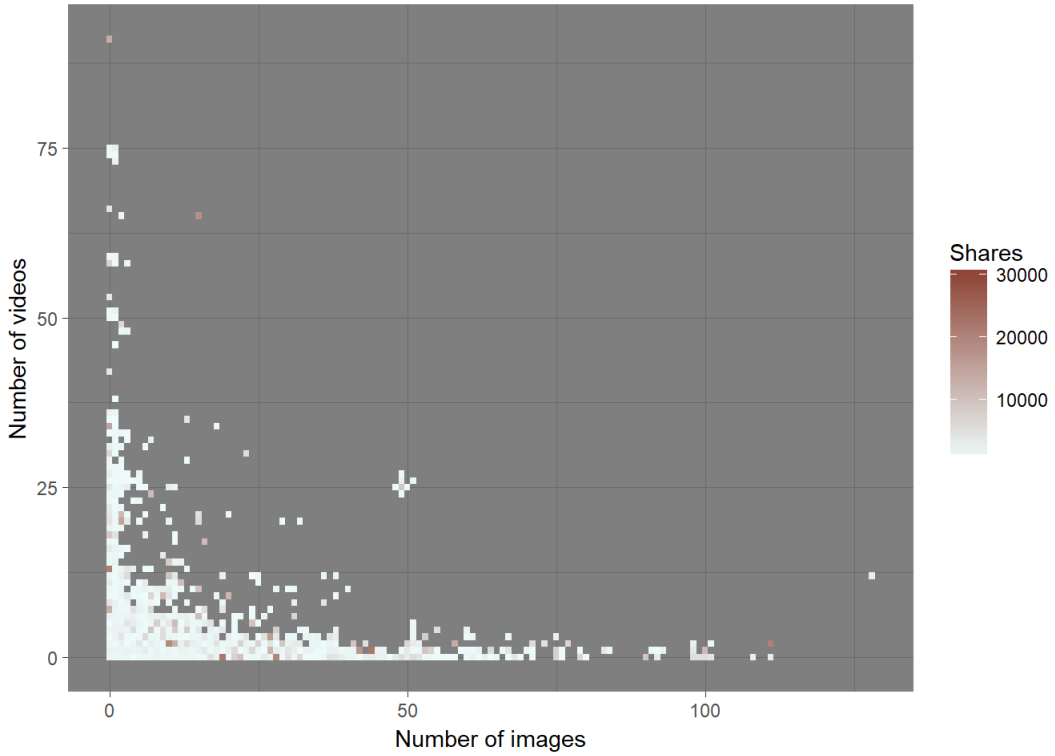
One more avenue of exploration is to look at the average positive and negative polarity of the article and the number of shares and see if anything interesting pops up. Let's look only at those articles which have > 10000 shares.



We see quite a few highy shared articles with 0 average negative polarity. Moreover, the bulk of these observations are in a tightly spanned area and there seems to be a negative correlation between average positive polarity and average negative polarity (as expected).

Effect of number of images and videos on number of shares

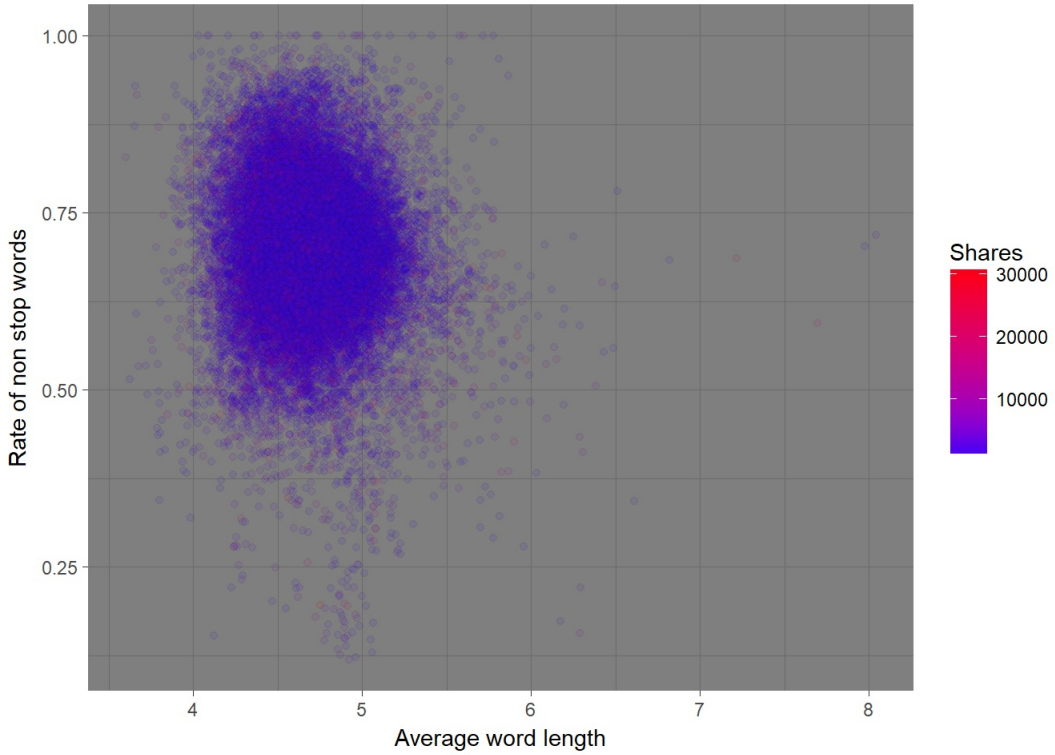
Let's now look at how the number of shares is influenced by number of videos and images. Again, we will only look at the bottom 99 percentile of the shares for this plot.



There are few articles with very high number of shares which have many images or videos. Otherwise, there is nothing interesting going on.

Effect of flow of the article on number of shares

The 'flow' of an article can be measured by the average word length and rate of non stop words in the content. A shorter average word length and a higher rate of non stop words in the content indicates a higher 'flow' to the article. Let's look at the effect of the (admiitedly subjective term) 'flow' on then number of shares.



There is a dense blob around which most of the entries are concentrated at. There is no interesting pattern or observation about the relation between average word length and rate of non stop words and number of shares that can be gleaned from the above scatterplot. We could possibly filter and only consider articles with more than 10000 shares, but going by the above scatter plot, we would expect to find a similar such blob around the same area.

Before we go further, let's just look at the number of articles which have more than 10000 shares to ensure that whatever conclusions or observations we made depended on enough number of data entries.

```
dim(news %>%
  filter (n_tokens_content > 0 & shares > 10000 & n_tokens_title > 0))
```

```
## [1] 2078   63
```

Looks okay.

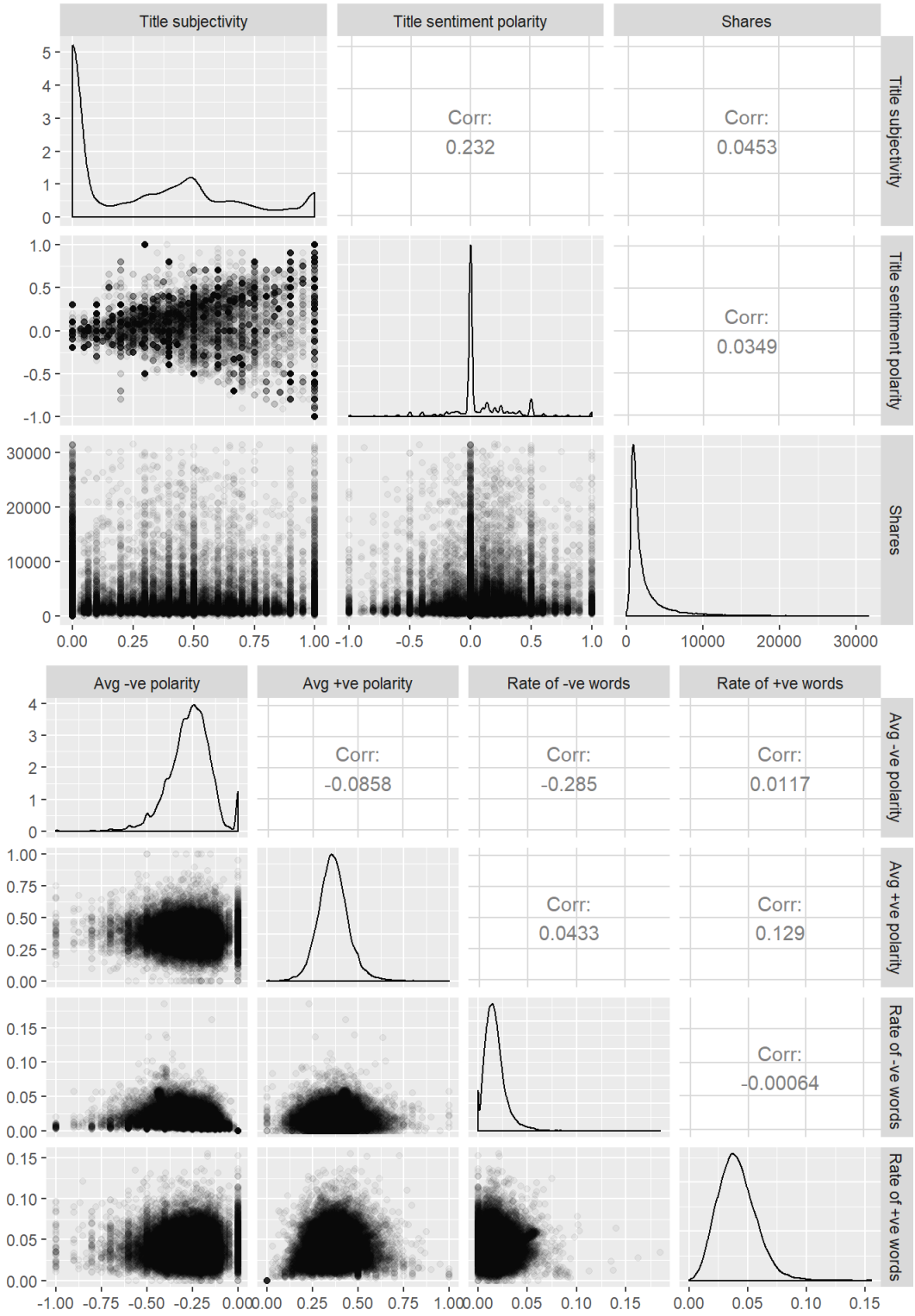
Bivariate Analysis

In this section, we chose to analyze pairs of variables and see if they had any effect on the number of shares. Variables which seemingly affect the popularity were chosen. We first looked at the tone and content of the articles and their titles and observed that there were lots of entries at fixed title subjectivity and title sentiment polarity. We then looked at only those articles which had more than 10000 shares and found that articles with higher shares tend to have either very subjective or very objective titles and tend to have neutral sentiment polarity. Further, we turned our attention to the average positive and negative polarity and number of shares and found out that there are a quite a few highly shared articles with exactly zero average negative polarity.

We then looked at the outlook of the article on the number of shares and found that for articles with very high number of shares, the global positive word rate tends to be higher than the global negative word rate. In other words, articles that have been shared heavily tend to be have more occurences of positive words. We then studied the effect of number of videos and images and the flow of article on the number of shares and found nothing very interesting.

Multivariate Plots Section

Based on what we found in the bivariate plots section, the relation between title sentiment, title subjectivity and number of shares is worth exploring. We could also look at the average positive and negative polarity, average rate of postive and negative words and see if there is any correlation between any two of the variables.



Multivariate Analysis

In this section, we futher explored the relationship between title subjectivity, title sentimentality and number of shares building on an observation in the section on Bivariate plots regarding these three variables. We can see the huge spike in the shares at title subjectivity = 0 and title subjectivity = 1 and title sentiment polarity = 0. We also see that articles tend to have a few fixed title subjectivity values as displayed in the title subjectivity vs title sentiment polarity . We can also see that the more subjective a title is, the greater the variation in the title sentiment polarity, which makes intuitive sense.

In the second plot, we choose to explore the relationship between variables that reflect the tone and subjectivity of the article. There are a couple of surprising observations here. First, rate of negative words and average negative polarity have a decent negative correlation. One would have expected that the more negatively polar an article is, the more the rate of negative words which is not to be. This can presumably be attributed to global polarity refering to a pessimistic outlook rather than frequency of negative words outlook. Second is the existence of substantially many articles with zero average negative polarity as can be seen from the plot of average negative polarity vs average positive polarity. Going by what we hyopthesized just earlier, this would imply that many articles on Mashable are unequivocally optimistic in their outlook.

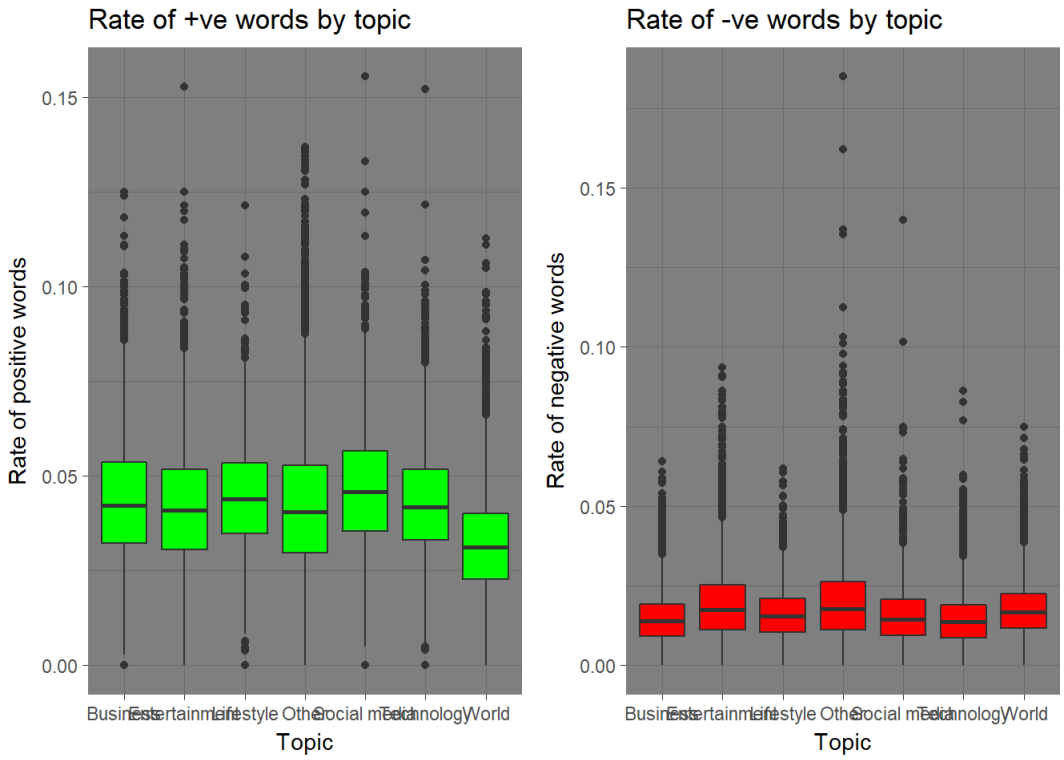
Final Plots and Summary

In this project, we explored the relationships between many variables and made some interesting observations. Before we proceed on to choosing three plots which would best present our most interesting findings, there are a couple of things that need to be clarified.

Notice that when plotting figures that involved number of shares, we only looked at the bottom 99 percentile of the shares. As we saw early on in the section on univariate plots, the distribution of shares is highly skewed and hence to get a better understanding of what is going on, we restricted ourselves to the bottom 99 percentile of shares. Second, when looking at the plots which involved title or content, we filtered out the articles which had no title (0 words in title) or no words (0 words in content). This was necessary as there were some articles with no words in title or no words in the content. Filtering out such articles gave us a truer picture of what was going on.

Plot One

Rate of postive and negative words split by topic



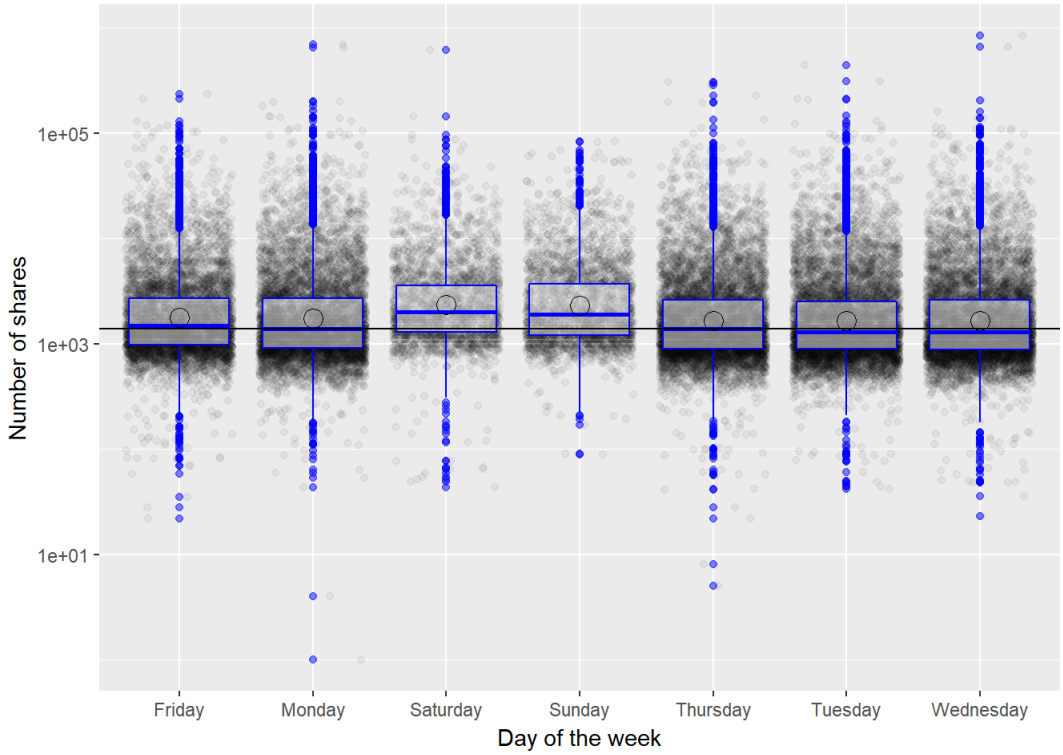

```
## # A tibble: 7 × 4
##            topic median_ptve_word_rate median_ntve_word_rate median_difference_rate
##            <chr>                  <dbl>                  <dbl>                   <dbl>
## 1       Business             0.04225352             0.01401869              0.02823483
## 2  Entertainment             0.04047619             0.01719902              0.02327717
## 3      Lifestyle             0.04377104             0.01538462              0.02838643
## 4          Other             0.03773585             0.01630435              0.02143150
## 5   Social media             0.04578755             0.01449275              0.03129479
## 6     Technology             0.04187315             0.01363249              0.02824066
## 7          World             0.03061224             0.01648352              0.01412873
```

Description one

This plot is interesting as it tells us that, regardless of the topic, postive words tend to occur more frequently than negative words, on an average. Moreover, the ‘World’ articles have the lowest postive word rate median and the difference between the postive and negative word rate median for ‘World’ articles is the the smallest. This plot gives more evidence to the hypothesis we had formed eariler that ‘World’ articles, on an average, tend to be more nuanced.

Plot two

Number of shares by day of publication



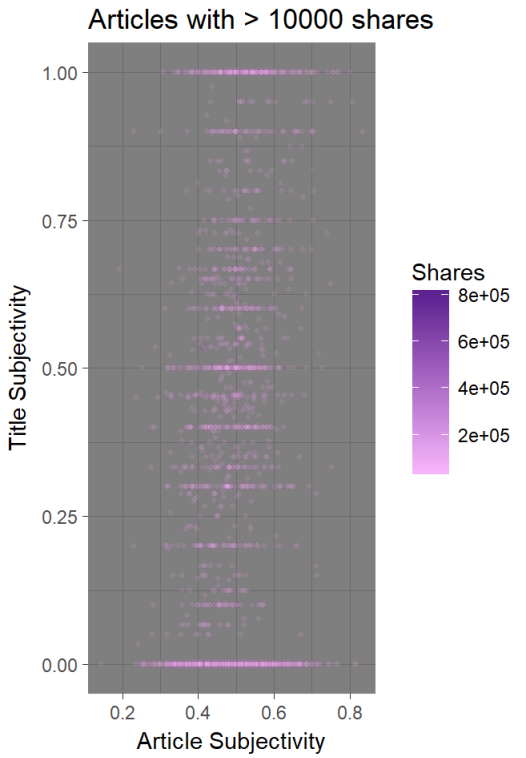

```
## # A tibble: 7 × 4
##   day_of_publication total median_shares avg_shares
##                <chr> <int>         <dbl>      <dbl>
## 1             Friday  5701          1500   3285.181
## 2             Monday  6661          1400   3647.026
## 3           Saturday  2453          2000   4078.185
## 4             Sunday  2737          1900   3746.741
## 5           Thursday  7267          1400   3178.599
## 6            Tuesday  7390          1300   3202.501
## 7          Wednesday  7435          1300   3303.405
```

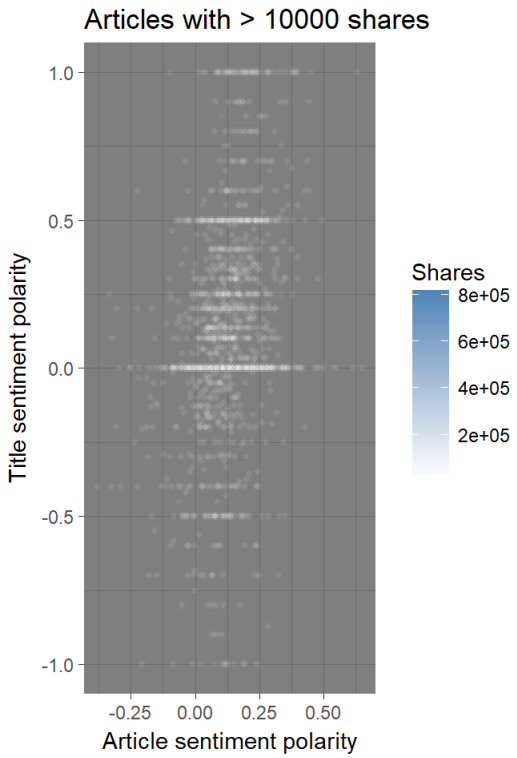
Description Two

The next plot we choose to display is a simple bar plot showing the median number of shares split by day. The horizontal line is the overall median of the number of shares. There are a couple of observations to be made. First, far fewer articles are actually published on weekends (which is expected). Second, inspite of this, the average number of shares (both mean and median) is higher for articles published on weekends. One can hypothesize that this happens as readers have more free time on weekends which means more time to read the articles (and hence share them.)

Plot Three

A look at effect of title and content among highy shared articles





Description Three

We see clusters of observations at title subjectivity = 0 and title subjectivity = 1 and title sentiment polarity around 0. For for articles which have high shares, the titles are either very subjective or very objective (1 or 0) and title sentiment polarity is around 0 (neutral). Article content does not seem to have any discernible effect on number of shares. This was most surprising as one would expect the content of an article to influence the popularity, but as seen in this plot, most popular articles have content ranging from subjective to objective and sentiment from positive to negative. (Note that we are not saying that title content causes popularity, just that for very popular articles, distinctly identifiable trends with respect to article titles emerge.)

Reflection

This finishes our exploratory data analysis on the online news popularity dataset. Choosing the correct variables for analysis amongst a wide variety of choices was bit of a struggle. In the end, we chose the variables which would presumably most influence the popularity and studied their interdependence. One minor concern was the mistake in data entry in some of the columns. But overall the data set was tidy and very comprehensive and we made some interesting observations, some of which confirmed our biases (or predispositions) and few of which refuted those. A very interesting project would be to use machine learning to predict the popularity of an article, given all the other variables.

Relevant Papers

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.