online-news-popularity

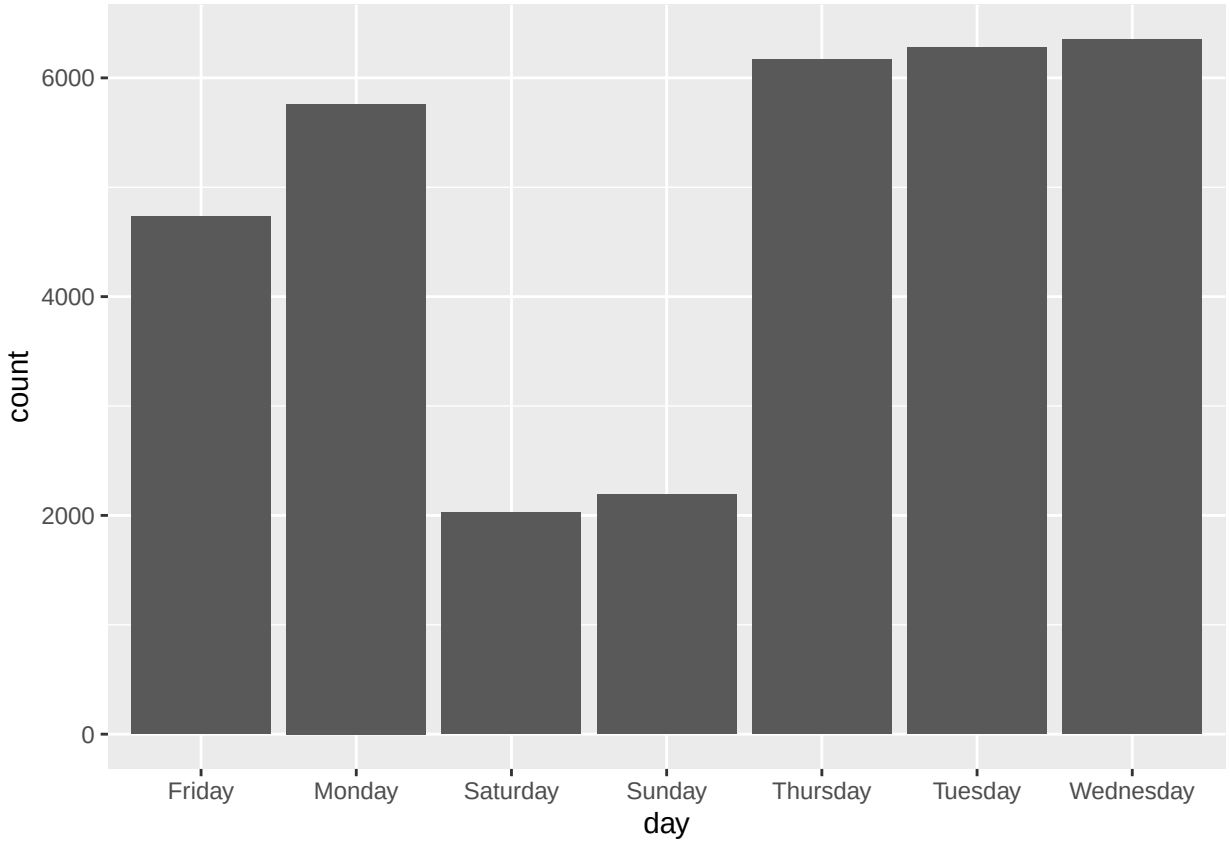
Sneh Gurdasani, Akshit Jain, Farhan Ansari, Sagar Singh

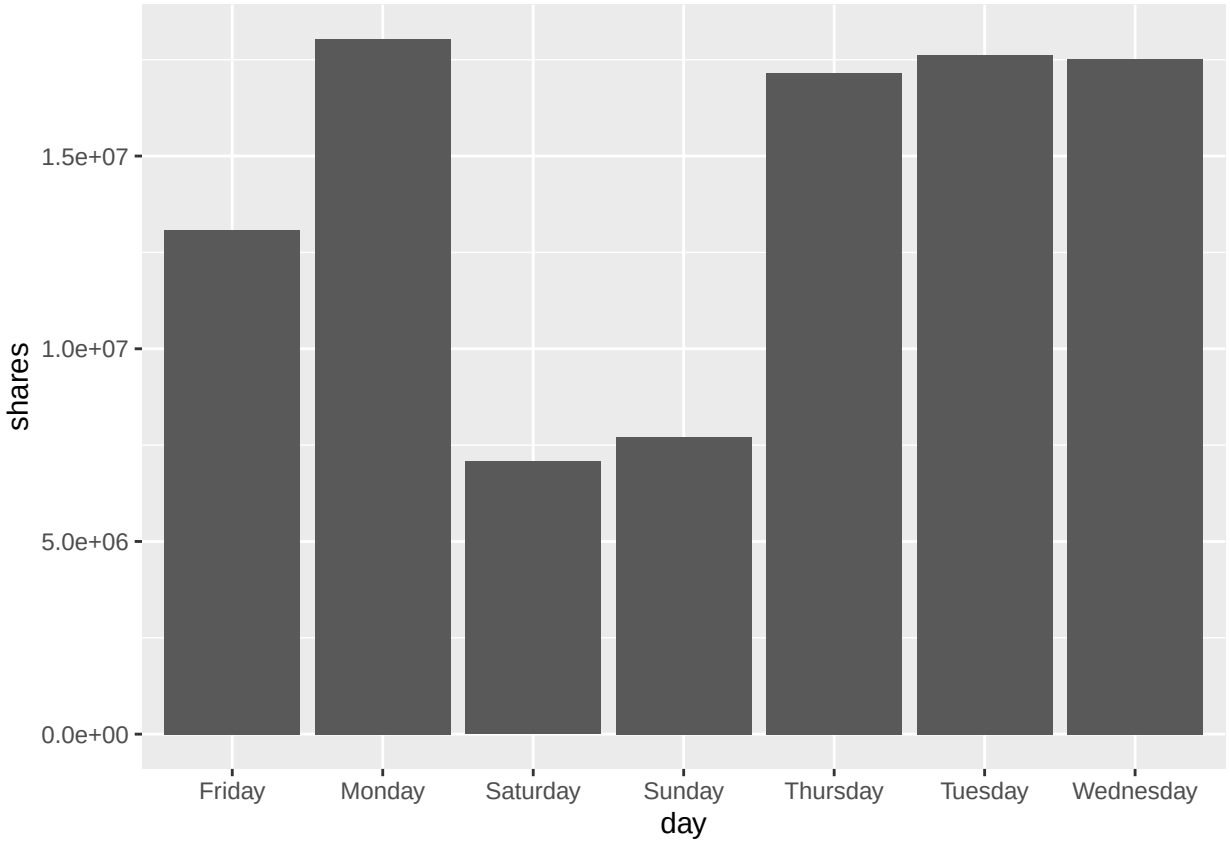
11/11/2019

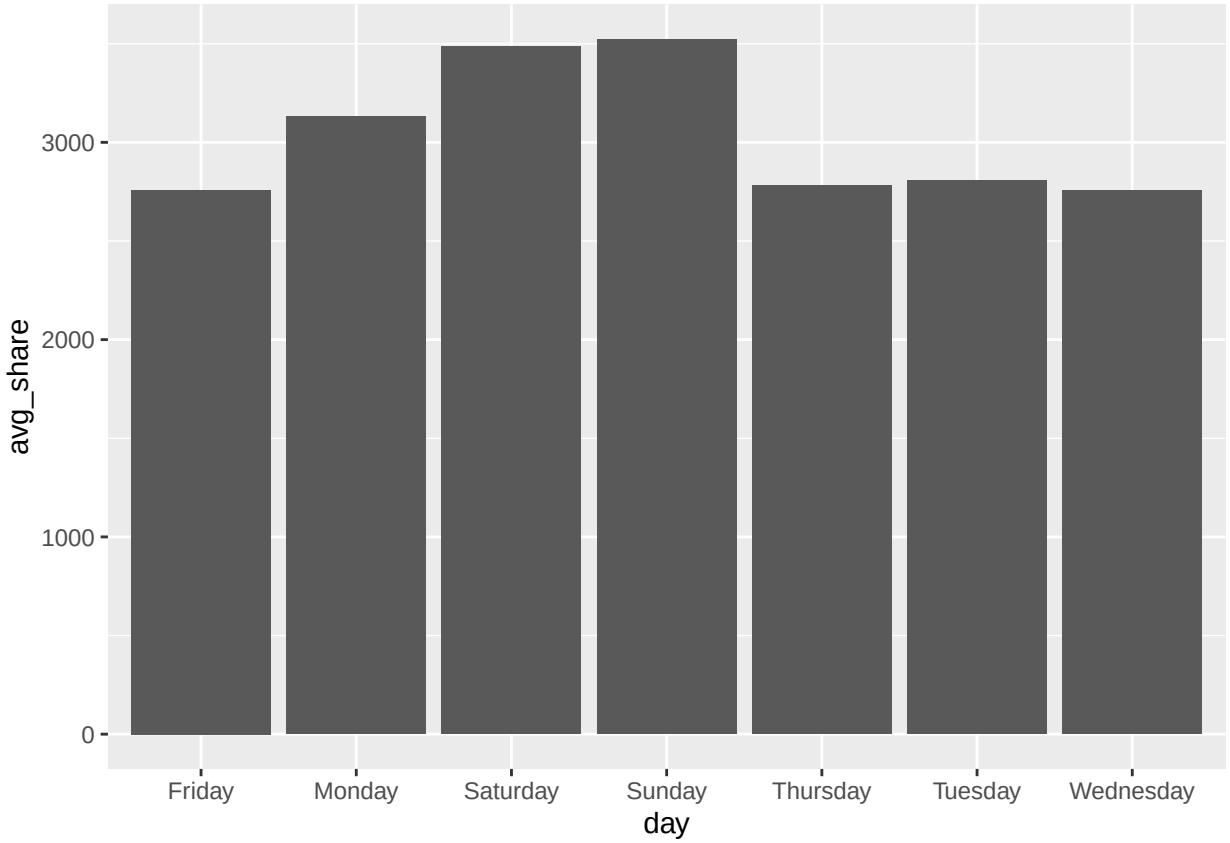
Load and Tidy Dataset

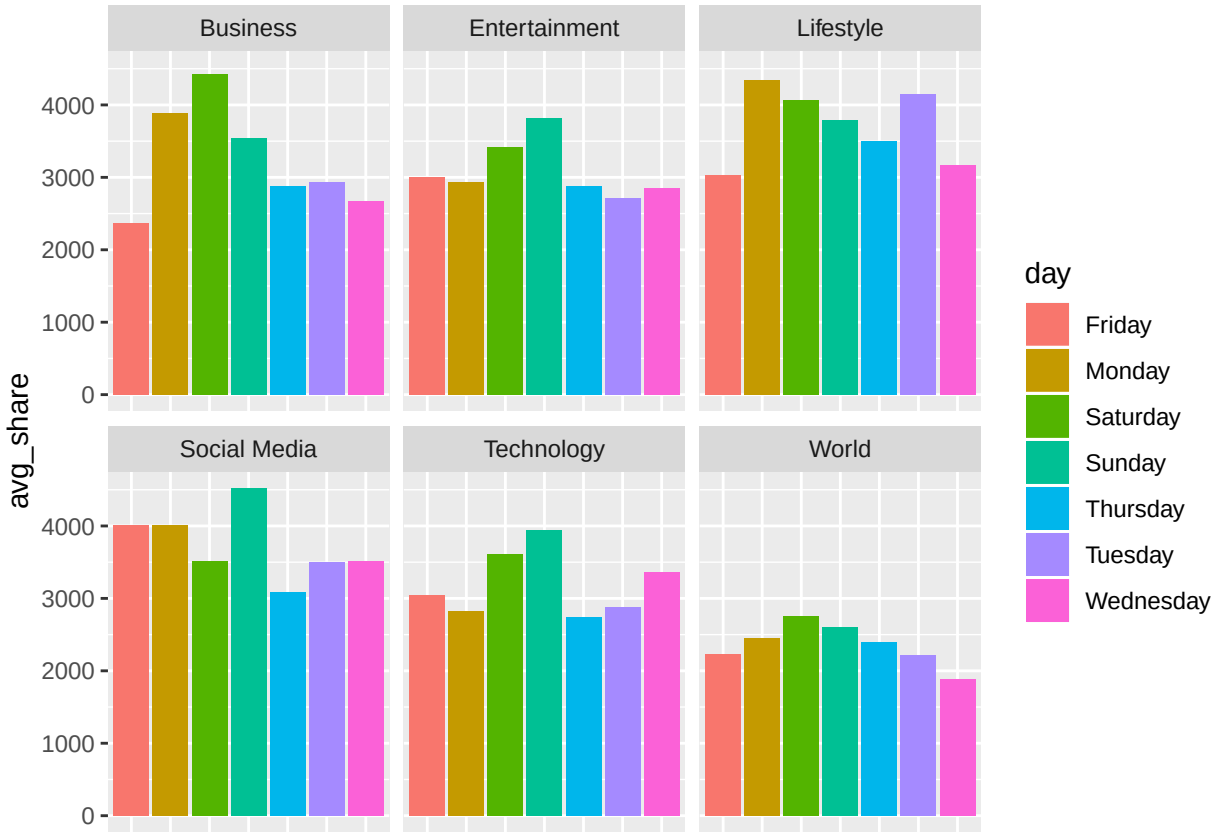
```
df <- read_csv('OnlineNewsPopularity/OnlineNewsPopularity.csv', col_types = cols())
list_of_dfs <- get_tidy_data(df)
cluster_numeric_df <- list_of_dfs[[1]]
classification_df <- list_of_dfs[[2]]
```

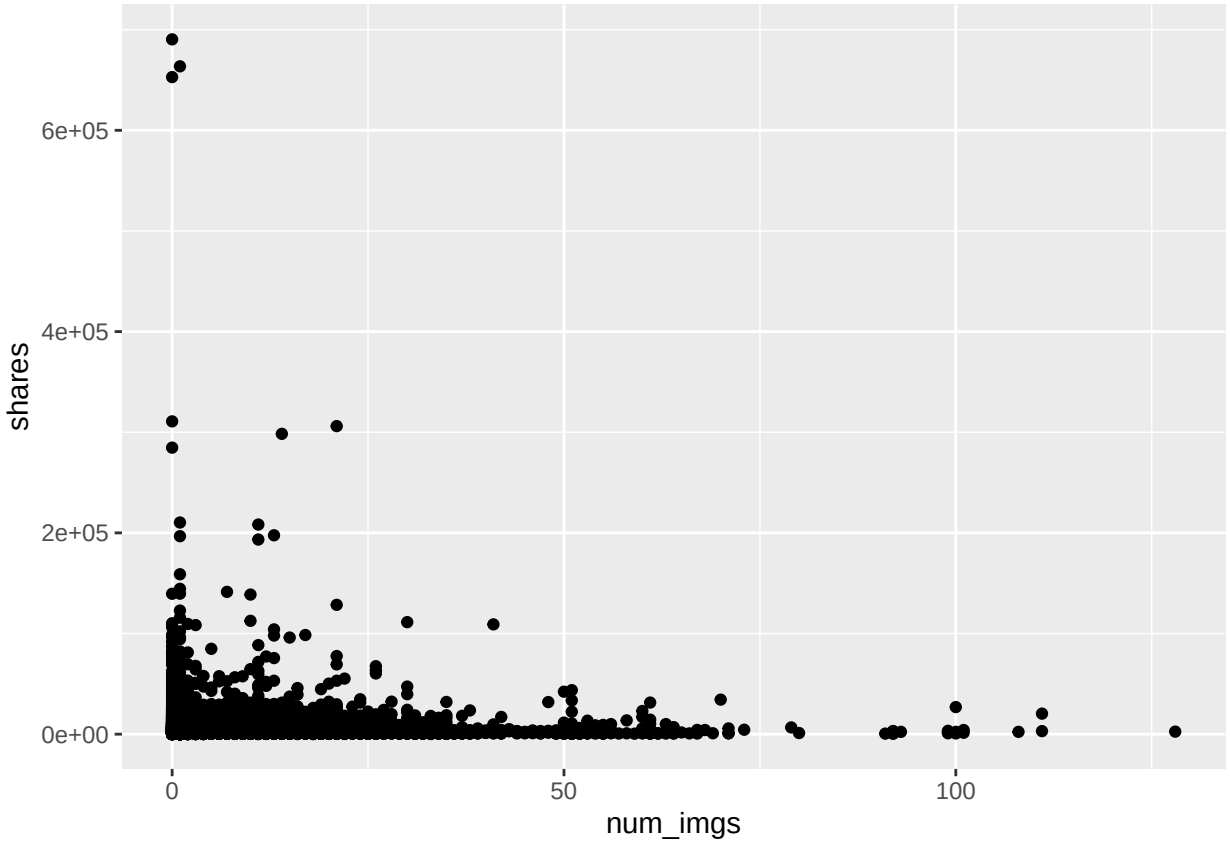
```
visualize_summary_plots(classification_df)
```



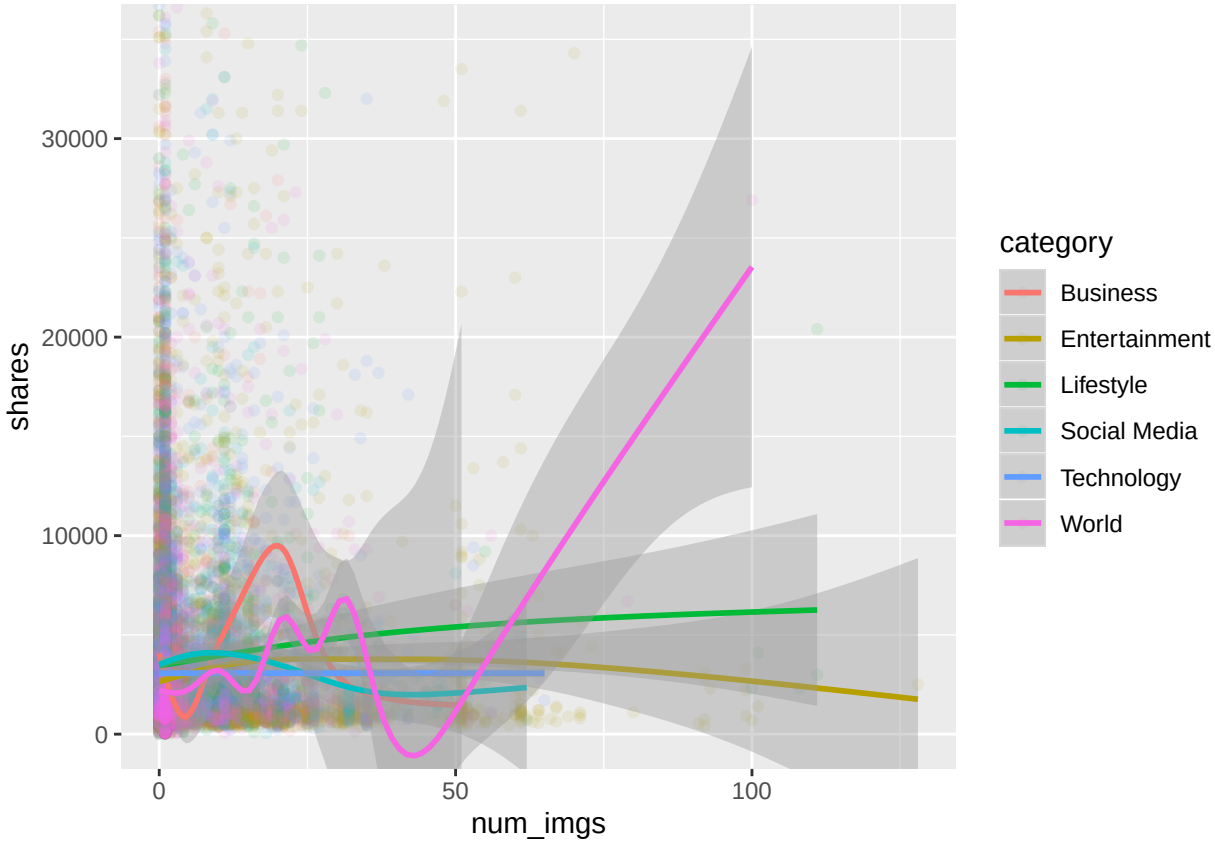

shares
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
day

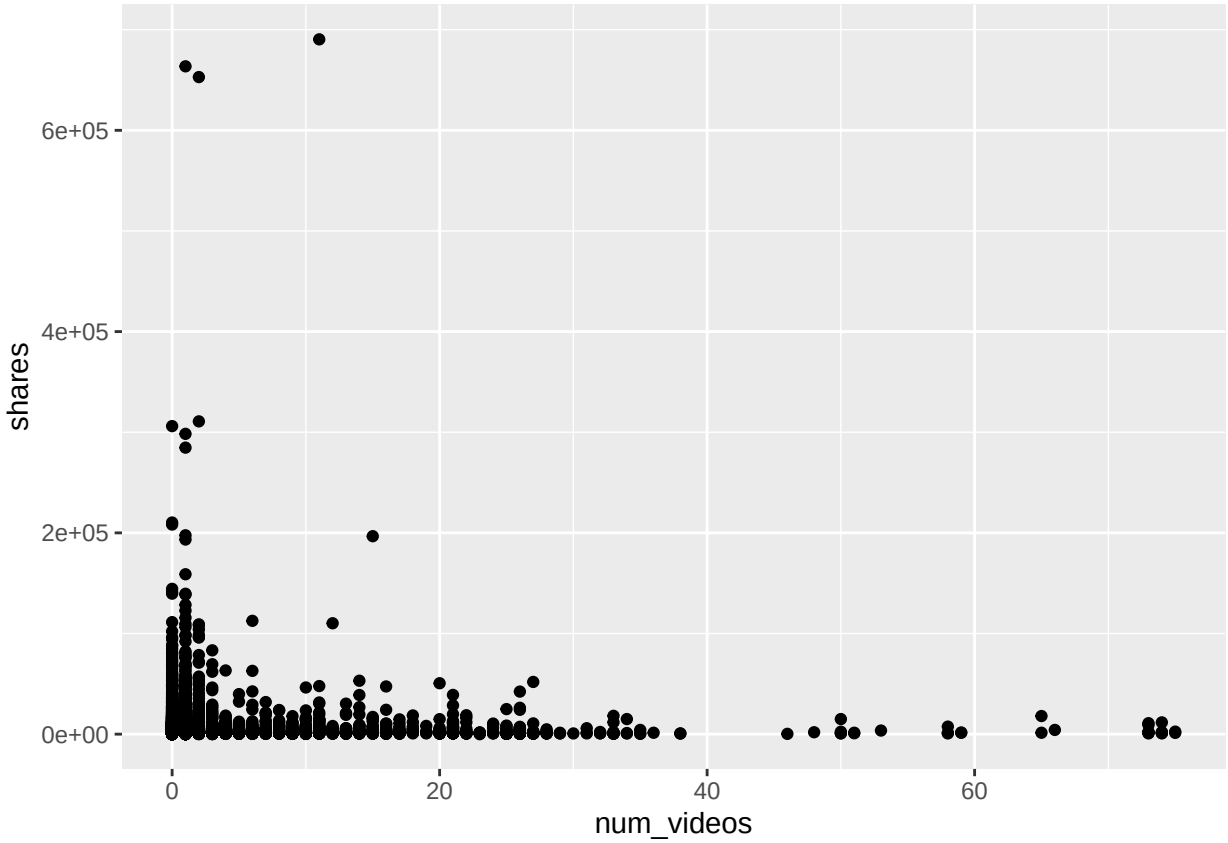

avg_share
3000
2000
1000
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
day


Business
Entertainment
Lifestyle
Social Media
Technology
World
avg_share
0
1000
2000
3000
4000
day
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

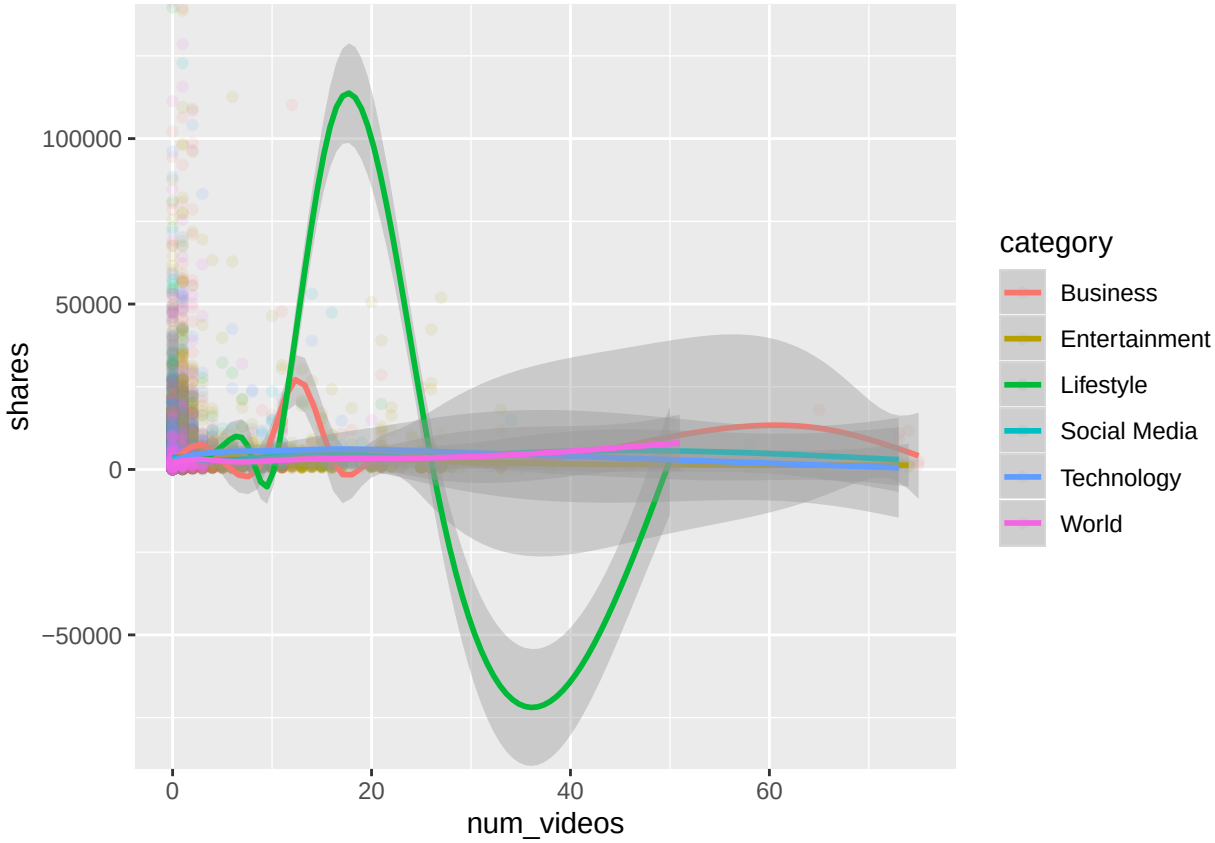


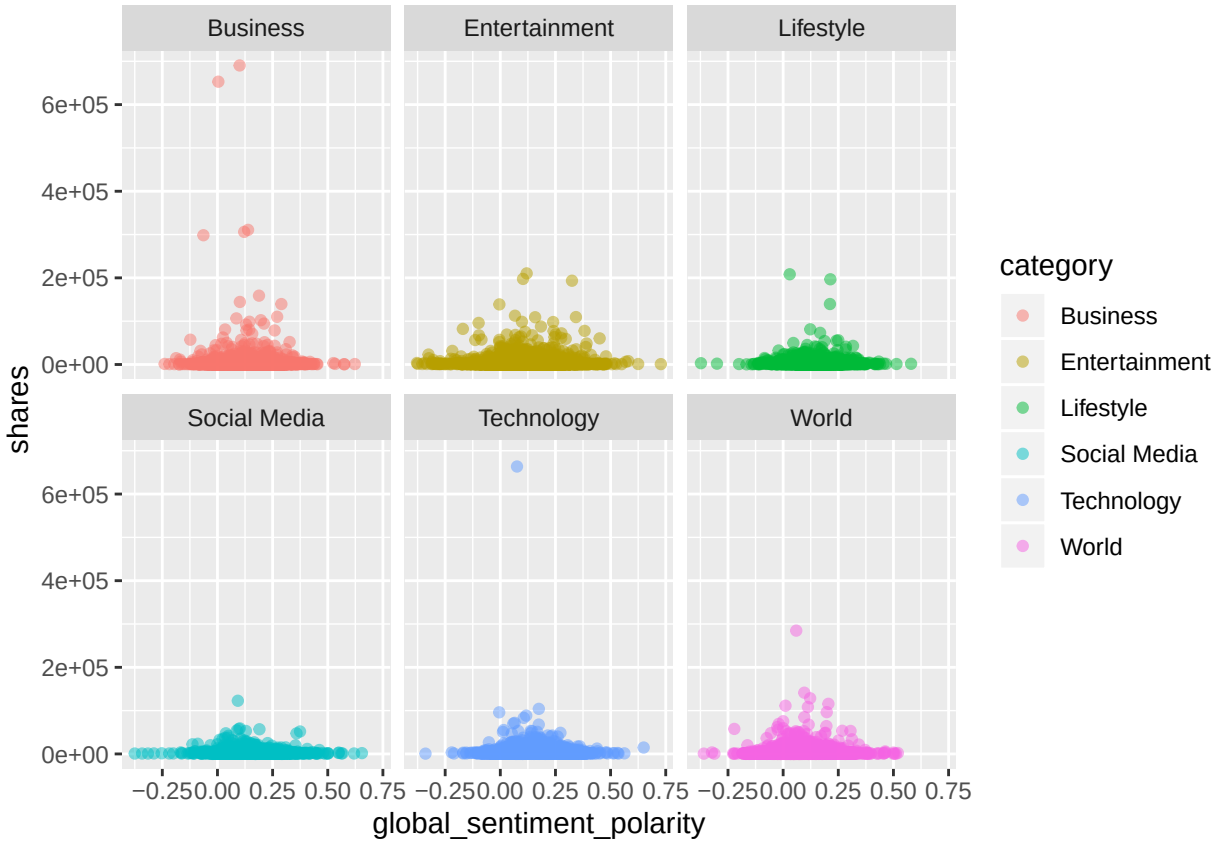

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

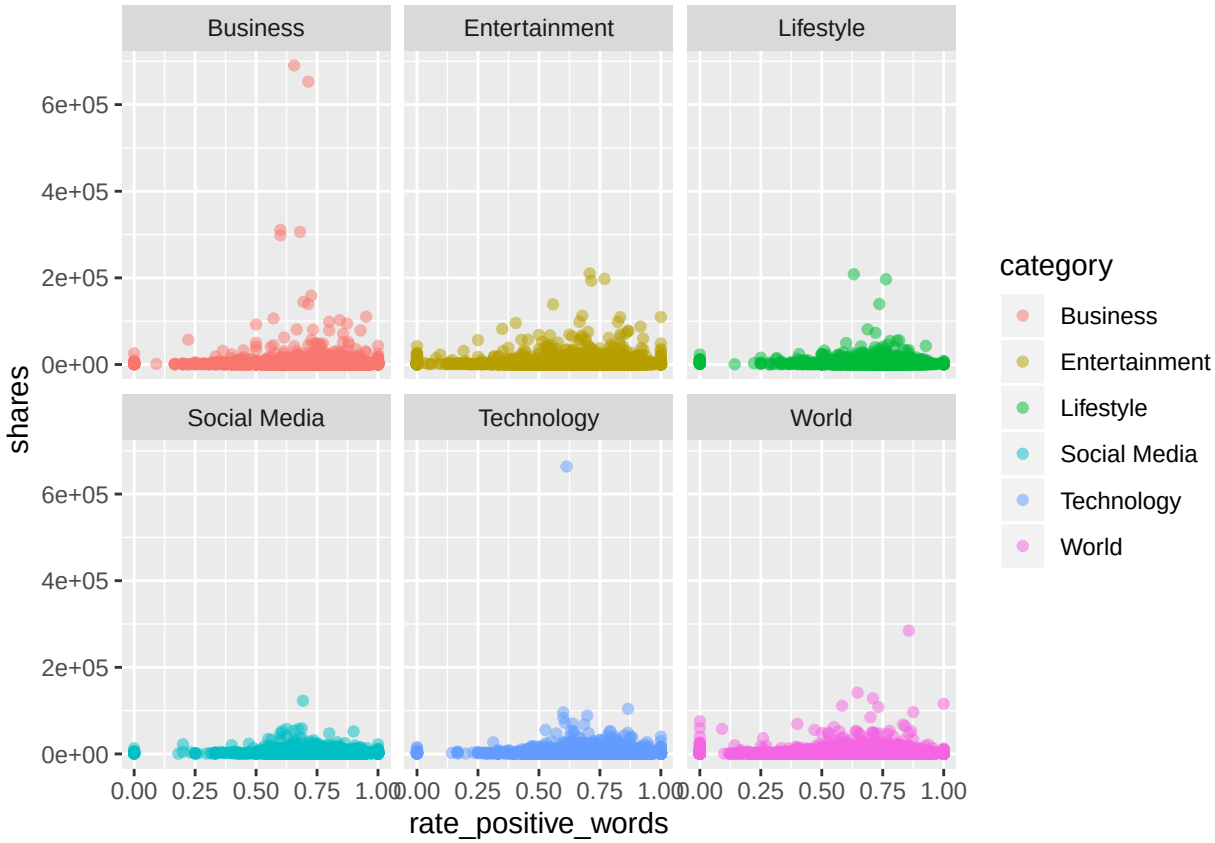

shares
30000
20000
10000
0
0
50
100
num_imgs
category
Business
Entertainment
Lifestyle
Social Media
Technology
World

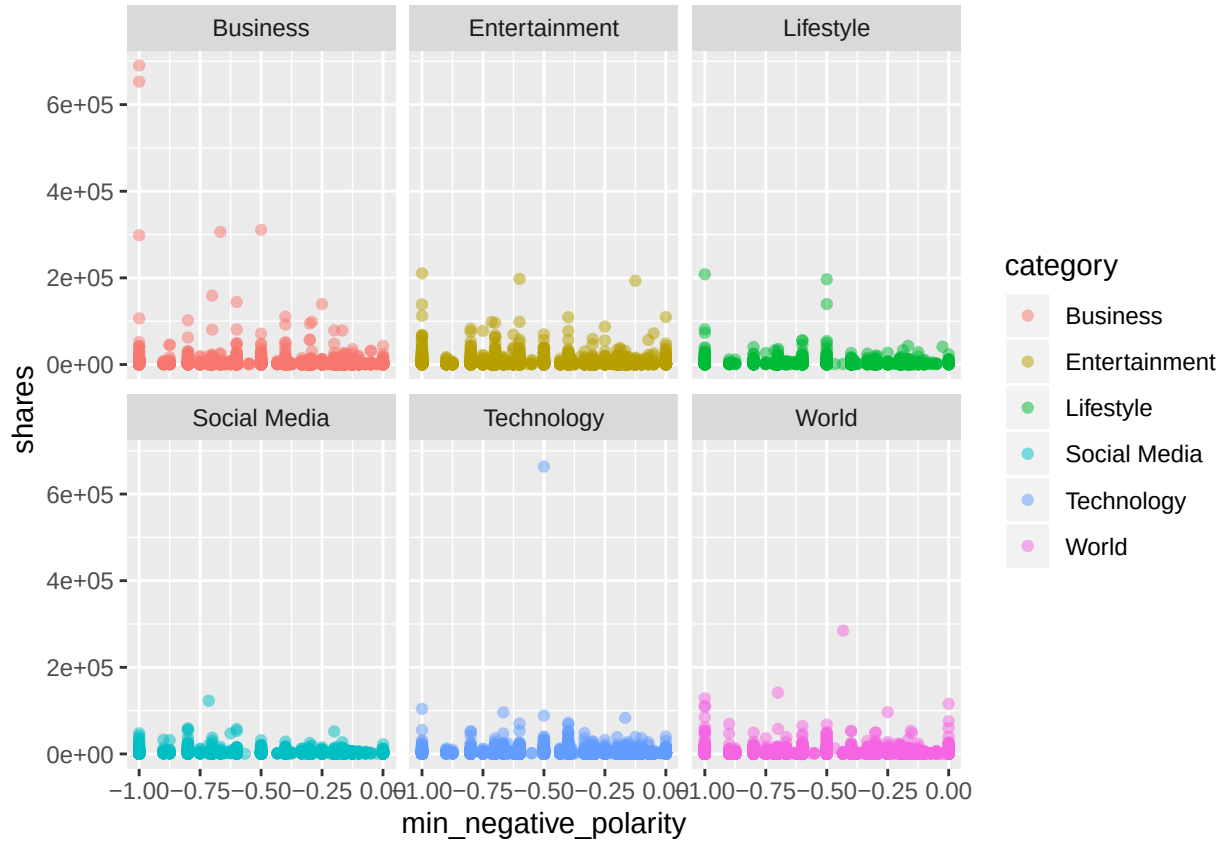



```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

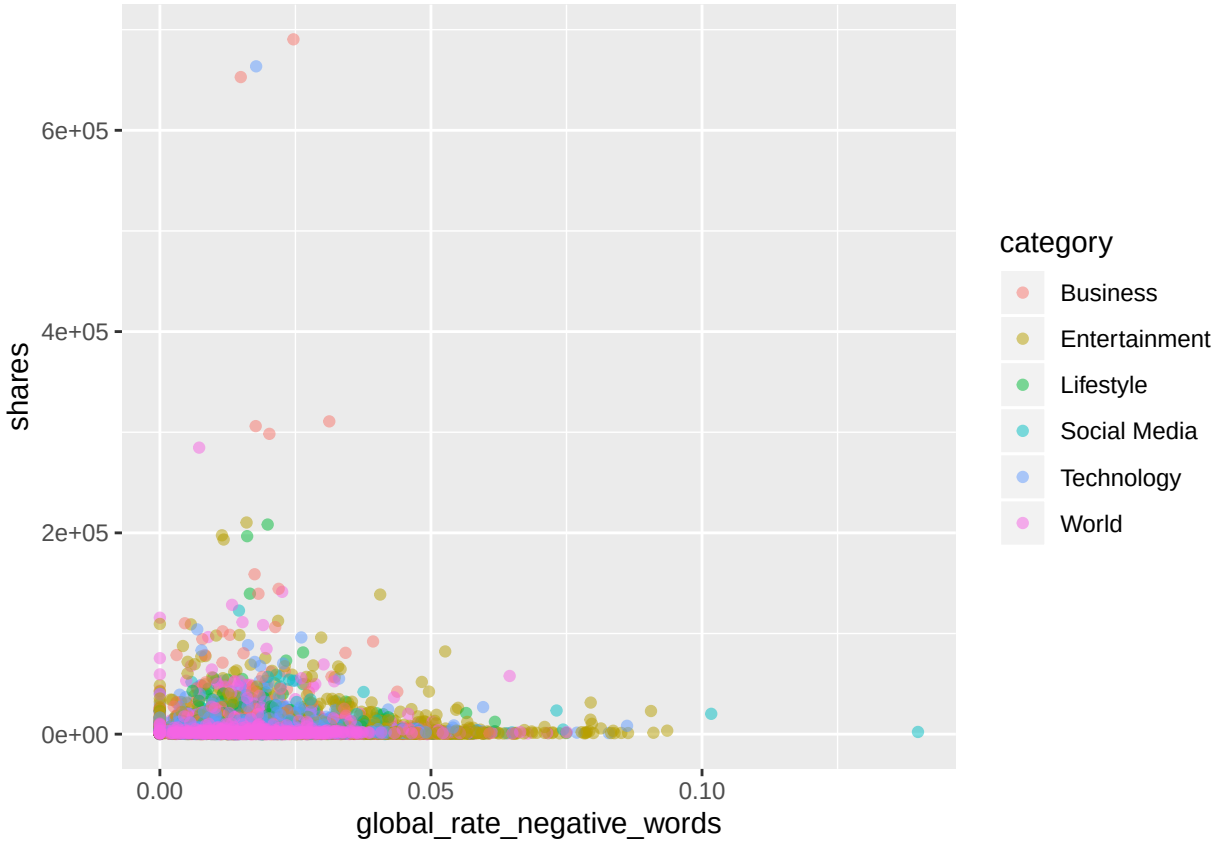

100000
50000
0
−50000
shares
0
20
40
60
num_videos
category
Business
Entertainment
Lifestyle
Social Media
Technology
World

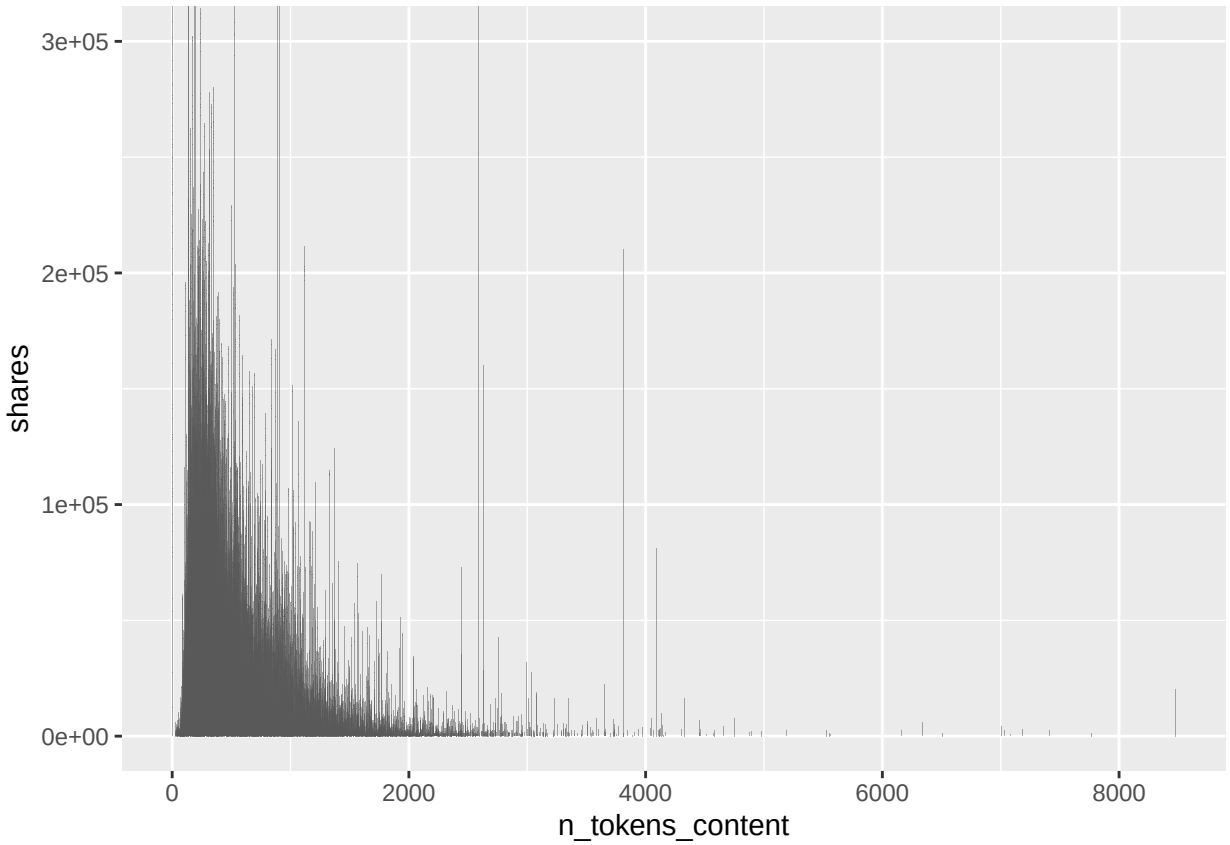

Business
Entertainment
Lifestyle
Social Media
Technology
World
shares
6e+05
4e+05
2e+05
0e+00
−0.25
0.00
0.25
0.50
0.75
global_sentiment_polarity
category
Business
Entertainment
Lifestyle
Social Media
Technology
World

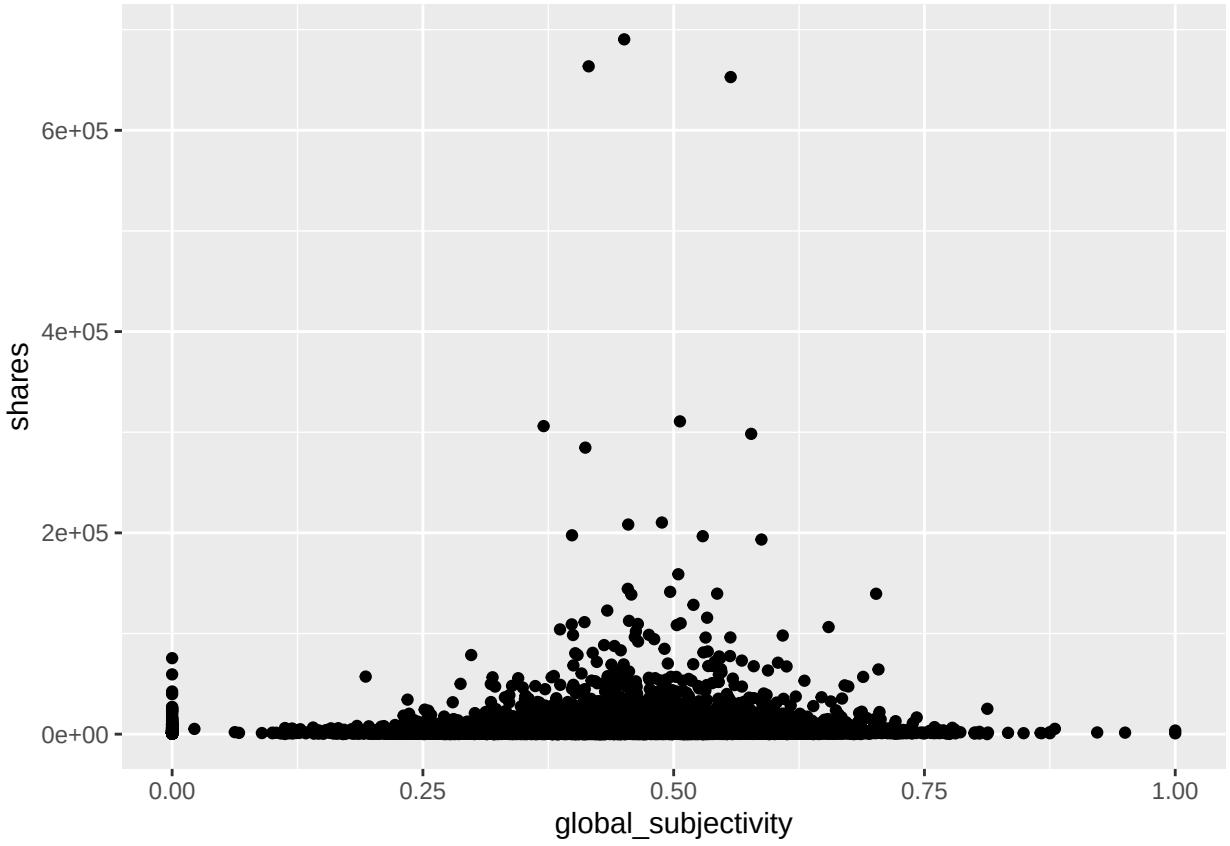

Business
Entertainment
Lifestyle
Social Media
Technology
World
shares
6e+05
4e+05
2e+05
0e+00
0.00
0.25
0.50
0.75
1.00
rate_positive_words
category
Business
Entertainment
Lifestyle
Social Media
Technology
World

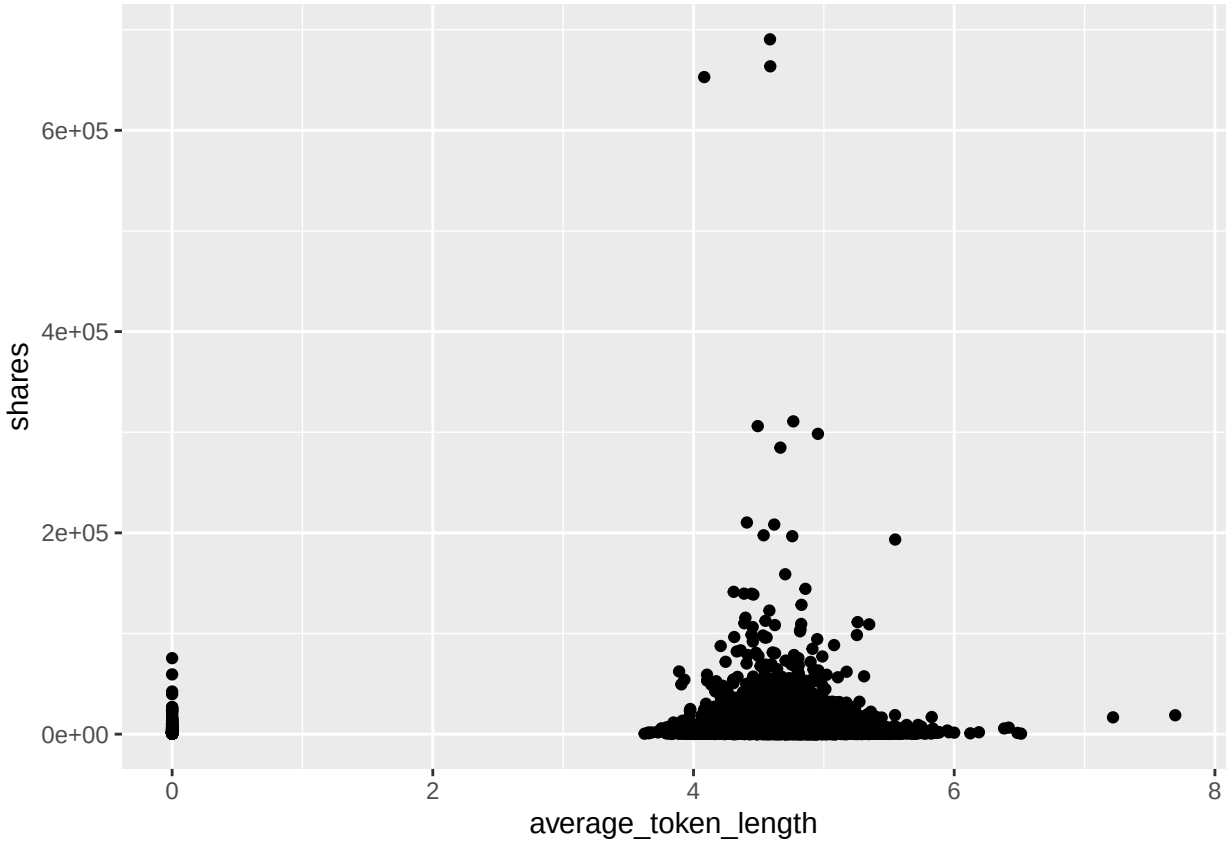



```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

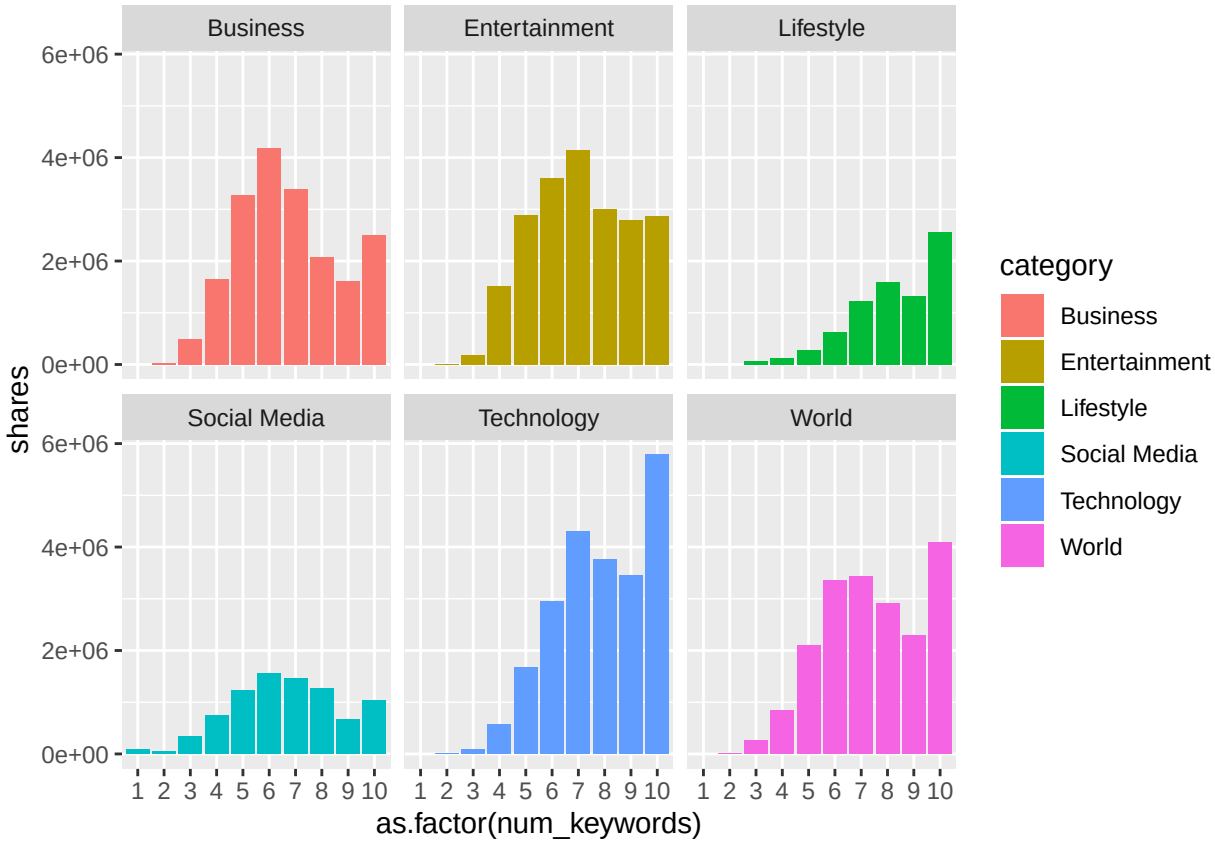

shares
6e+05
4e+05
2e+05
0e+00
0.00
0.05
0.10
global_rate_negative_words
category
Business
Entertainment
Lifestyle
Social Media
Technology
World

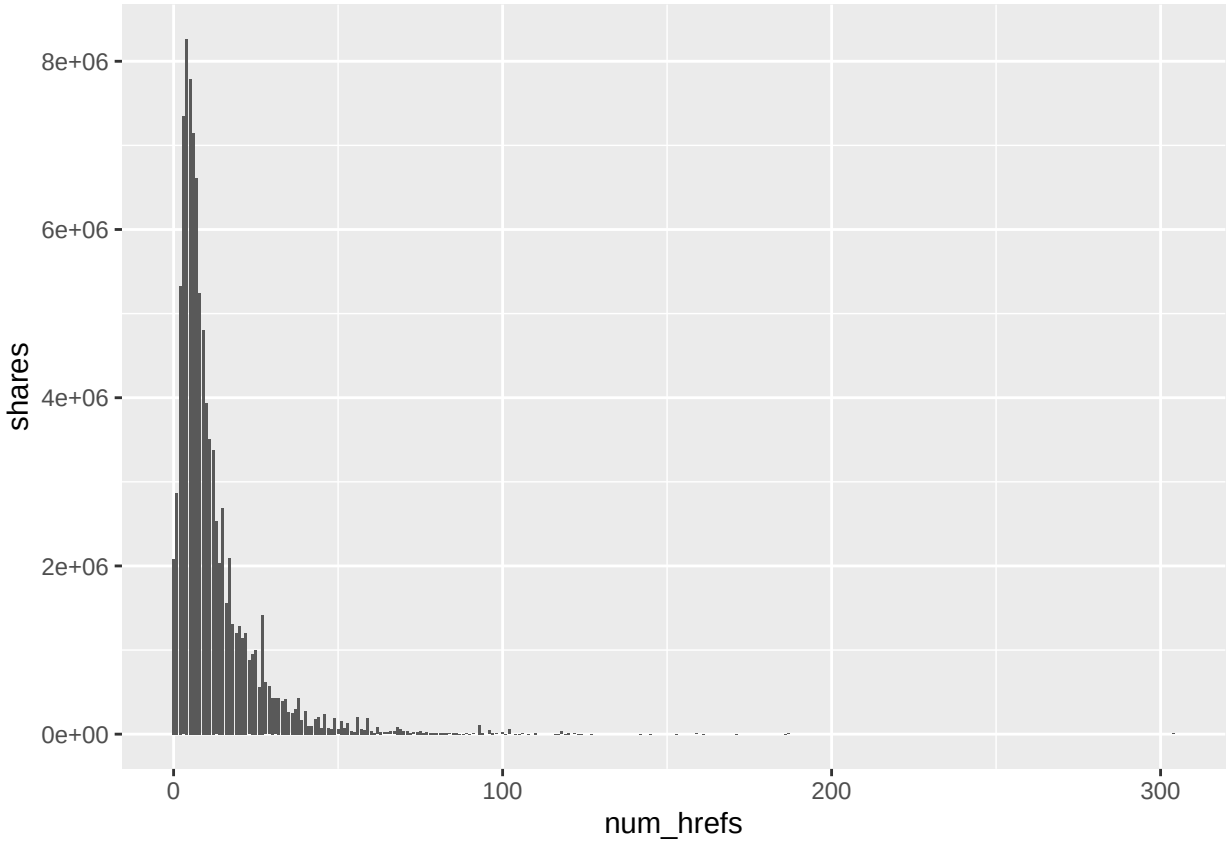

3e+05
2e+05
1e+05
0e+00
shares
0
2000
4000
6000
8000
n_tokens_content

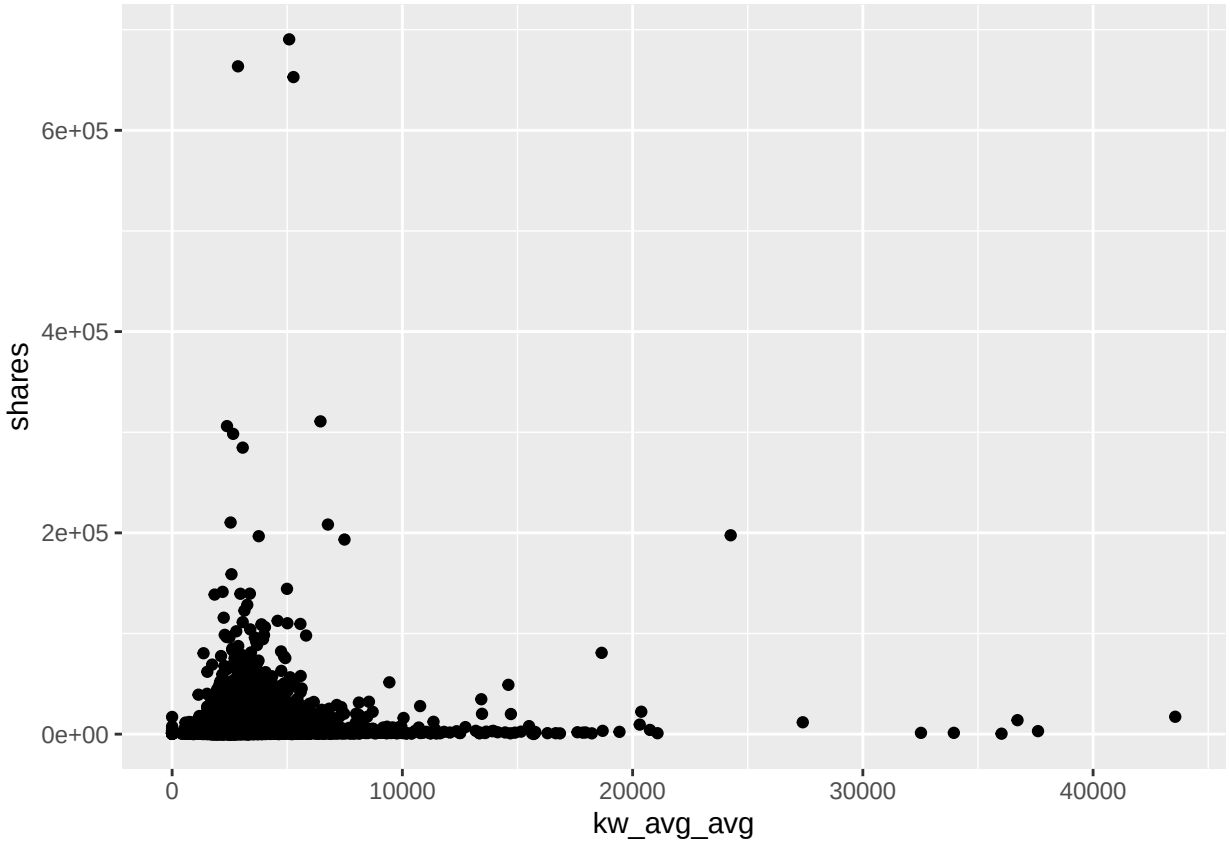

6e+05
4e+05
2e+05
0e+00
shares
0.00
0.25
0.50
0.75
1.00
global_subjectivity




```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


Business
Entertainment
Lifestyle
Social Media
Technology
World
shares
6e+06
4e+06
2e+06
0e+00
1 2 3 4 5 6 7 8 9 10
as.factor(num_keywords)
category
Business
Entertainment
Lifestyle
Social Media
Technology
World


8e+06
6e+06
4e+06
2e+06
0e+00
shares
0
100
200
300
num_hrefs




```
correlation_matrix(cluster_numeric_df)
```

```
## Selecting by shares
```

```
##              correlation_with     shares
## 1                      shares 1.00000000
## 2                  kw_avg_avg 0.10002583
## 3                  kw_max_avg 0.06586856
## 4  self_reference_avg_sharess 0.06145828
## 5   self_reference_min_shares 0.05825928
## 6   self_reference_max_shares 0.04895216
## 7                  kw_avg_min 0.04316214
## 8                  kw_max_min 0.04307623
## 9                   num_hrefs 0.04190182
## 10        global_subjectivity 0.03510054
## 11                     LDA_03 0.03288953
## 12                 kw_min_avg 0.02912447
## 13           n_tokens_content 0.02912325
## 14                   num_imgs 0.02877061
## 15               num_keywords 0.02554536
```

```
prepare_data_classification(classification_df)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     930    1400    2929    2500  690400
```