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_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, and data_channel_is_world – Indicator variables for the article belonging to different categories or "data Channels", we can merge into one categorical variable.

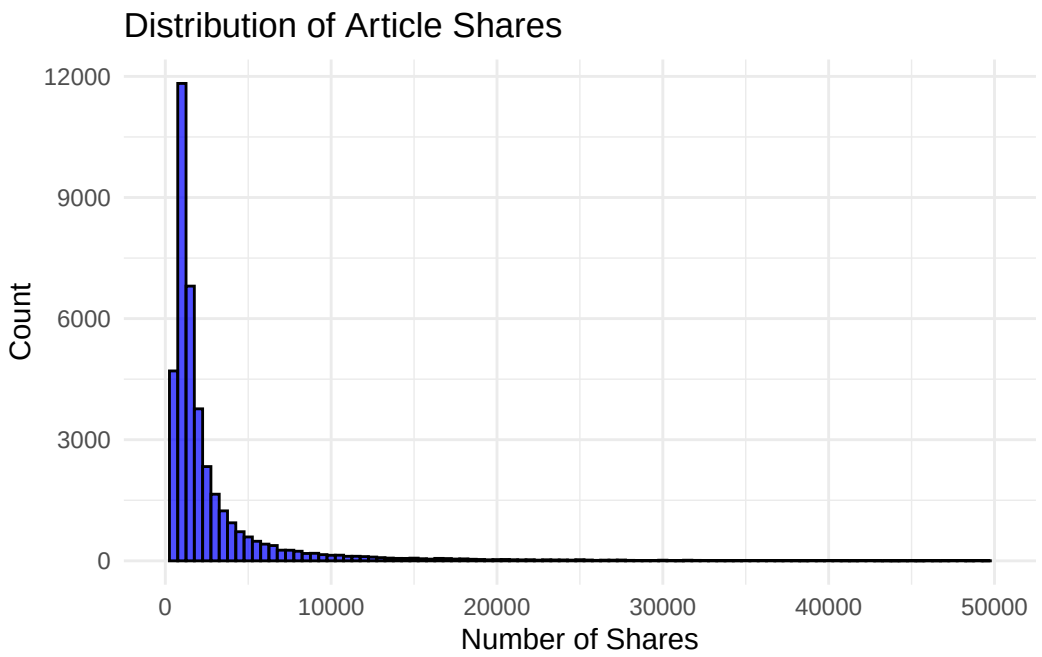
Weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday,weekday_is_saturday, weekday_is_sunday – These are indicator variables for what day the article was published, which we can merge into a categorical variable of Day_published

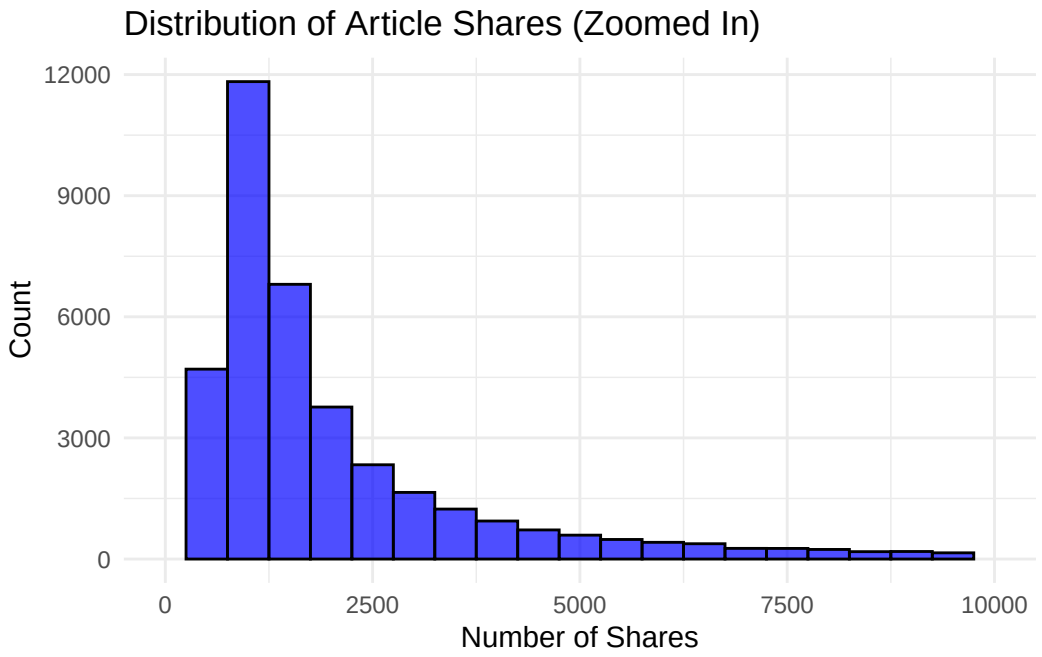
Data Cleaning

First, to we will convert the variables for day of the week and article topic into categorical variables for easier visualization.

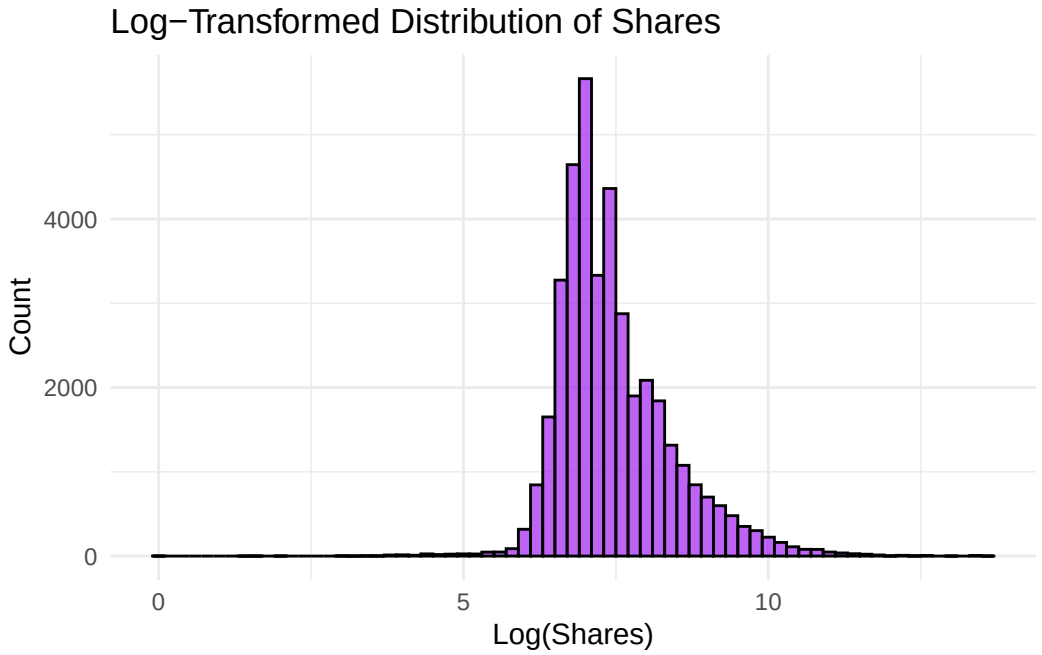
```
# A tibble: 6 x 63
  url                   timedelta n_tokens_title n_tokens_content n_unique_tokens
  <chr>                     <dbl>          <dbl>            <dbl>           <dbl>
1 http://mashable.com~       731             12              219           0.664
2 http://mashable.com~       731              9              255           0.605
3 http://mashable.com~       731              9              211           0.575
4 http://mashable.com~       731              9              531           0.504
5 http://mashable.com~       731             13             1072           0.416
6 http://mashable.com~       731             10              370           0.560
# i 58 more variables: n_non_stop_words <dbl>, n_non_stop_unique_tokens <dbl>,
#   num_hrefs <dbl>, num_self_hrefs <dbl>, num_imgs <dbl>, num_videos <dbl>,
#   average_token_length <dbl>, num_keywords <dbl>,
#   data_channel_is_lifestyle <dbl>, data_channel_is_entertainment <dbl>,
#   data_channel_is_bus <dbl>, data_channel_is_socmed <dbl>,
#   data_channel_is_tech <dbl>, data_channel_is_world <dbl>, kw_min_min <dbl>,
#   kw_max_min <dbl>, kw_avg_min <dbl>, kw_min_max <dbl>, kw_max_max <dbl>, ...
```

Response Variable EDA

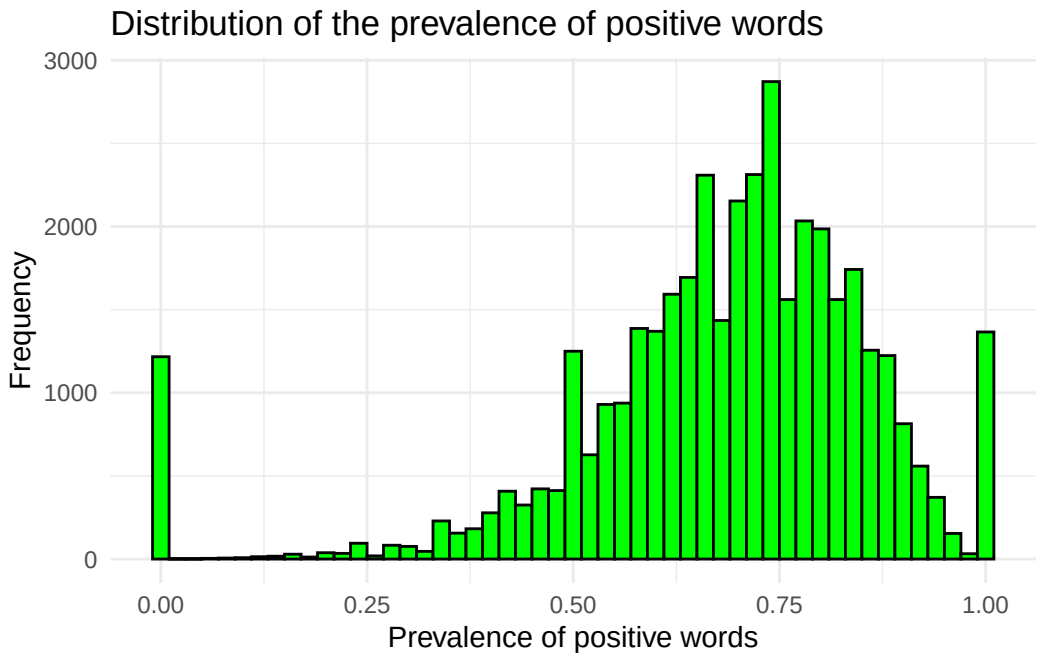


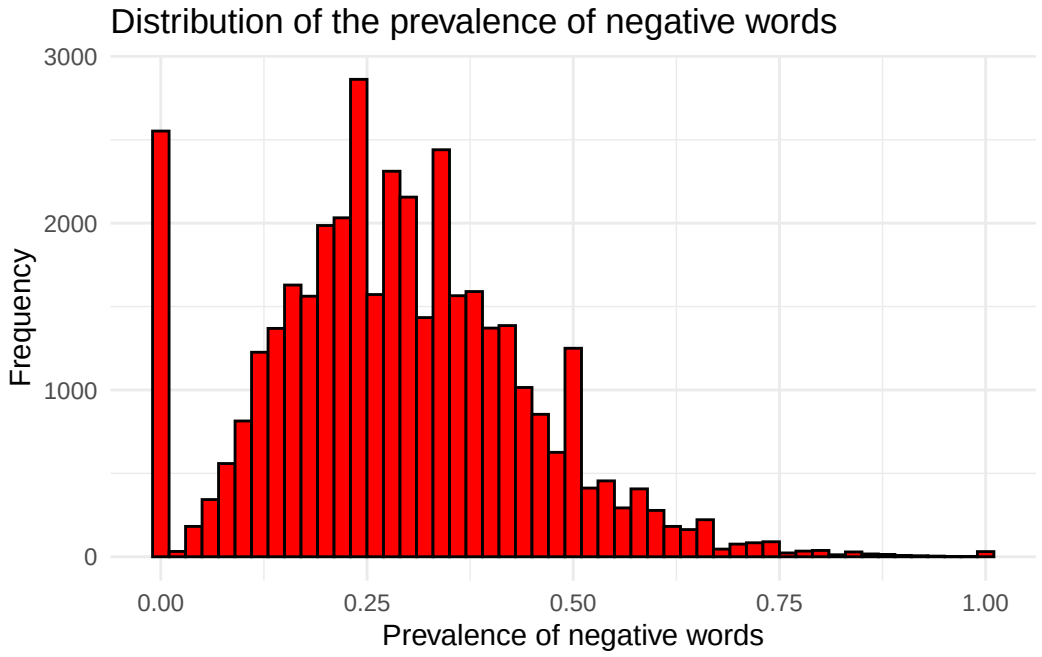


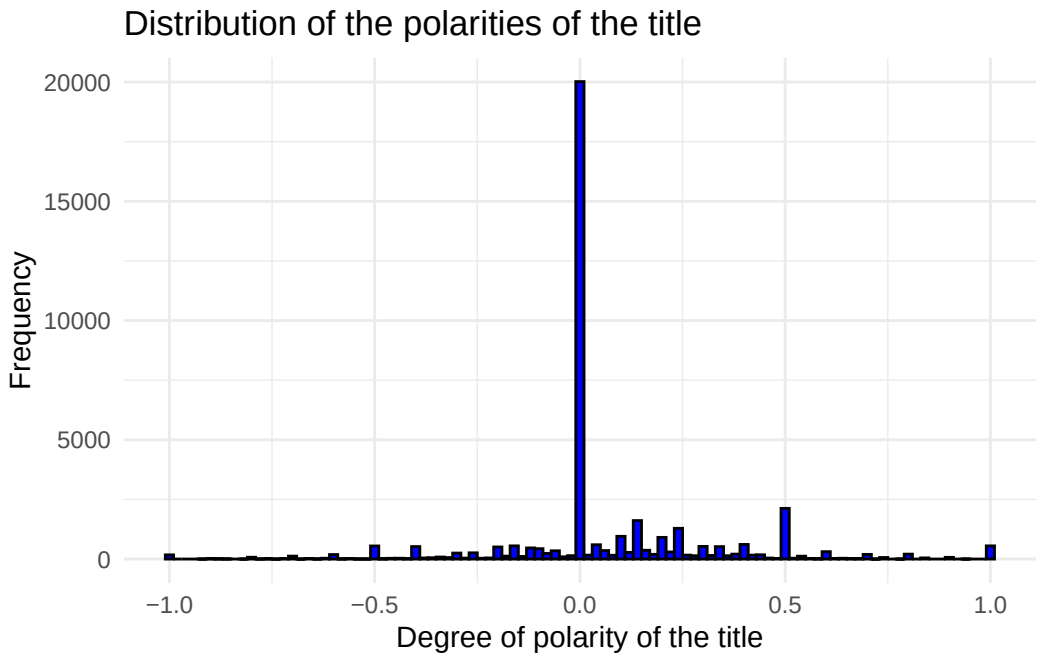
Due to the right skewed nature of our data set, we decided to use a log transformation so that the response variable would be normally distributed.

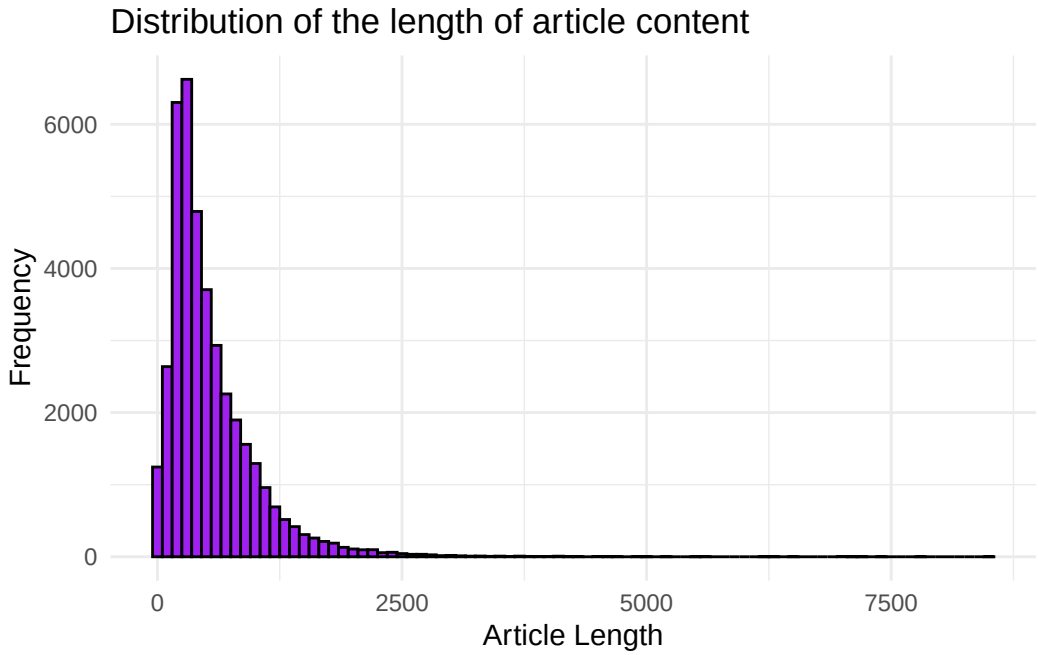


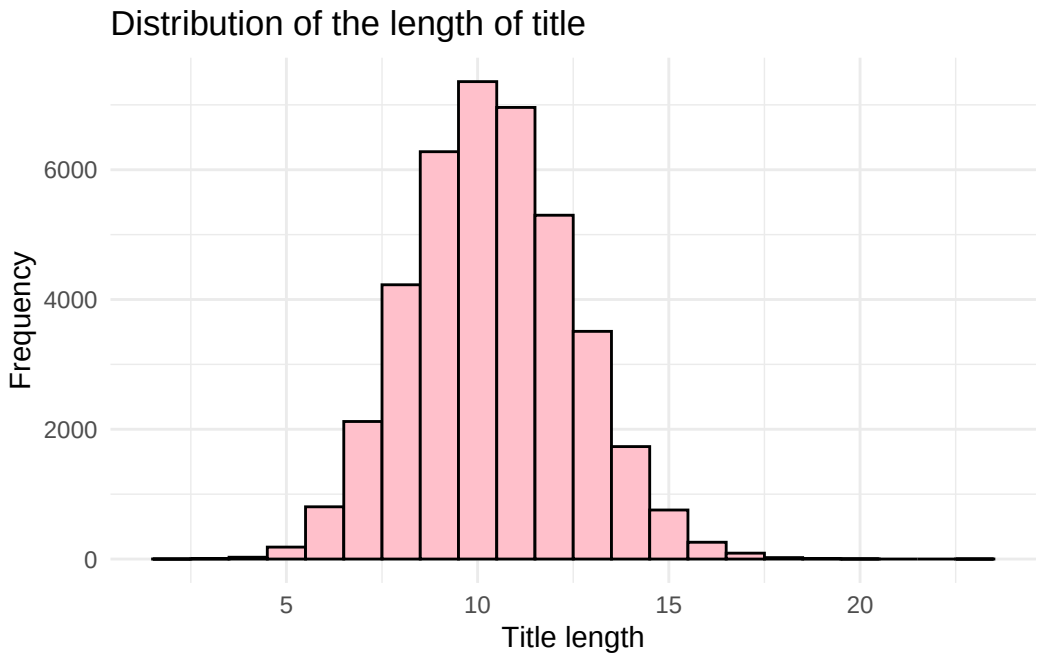
Predictor Variable EDA


Distribution of the prevalence of positive words
Frequency
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Prevalence of positive words


Distribution of the prevalence of negative words
Frequency
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Prevalence of negative words

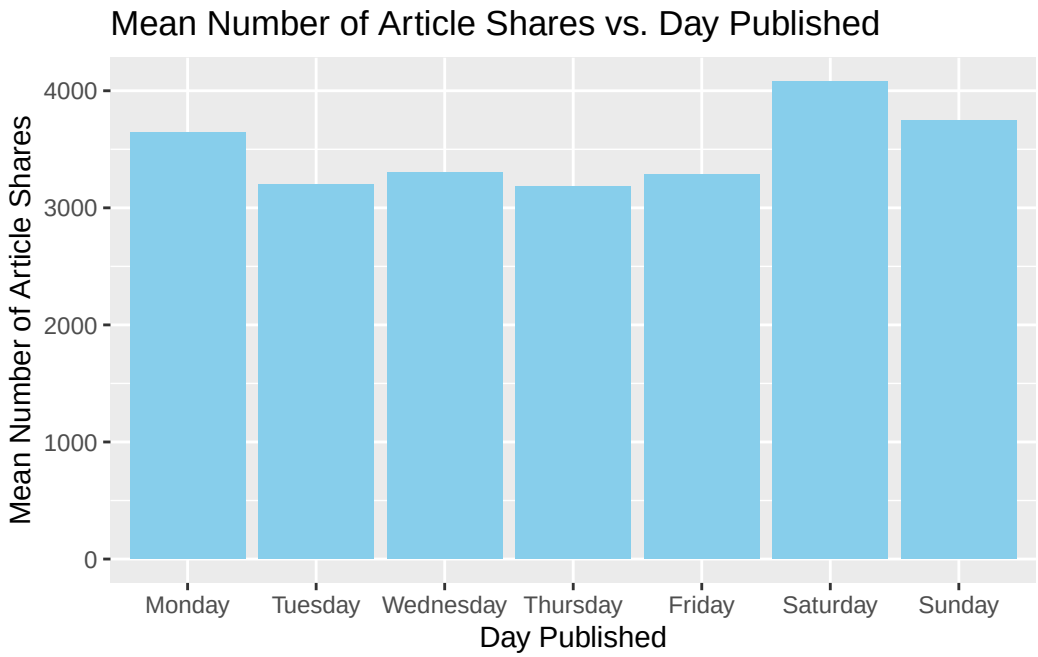

Distribution of the polarities of the title
Frequency
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0
Degree of polarity of the title


Distribution of the length of article content
Frequency
6000
4000
2000
0
0
2500
5000
7500
Article Length



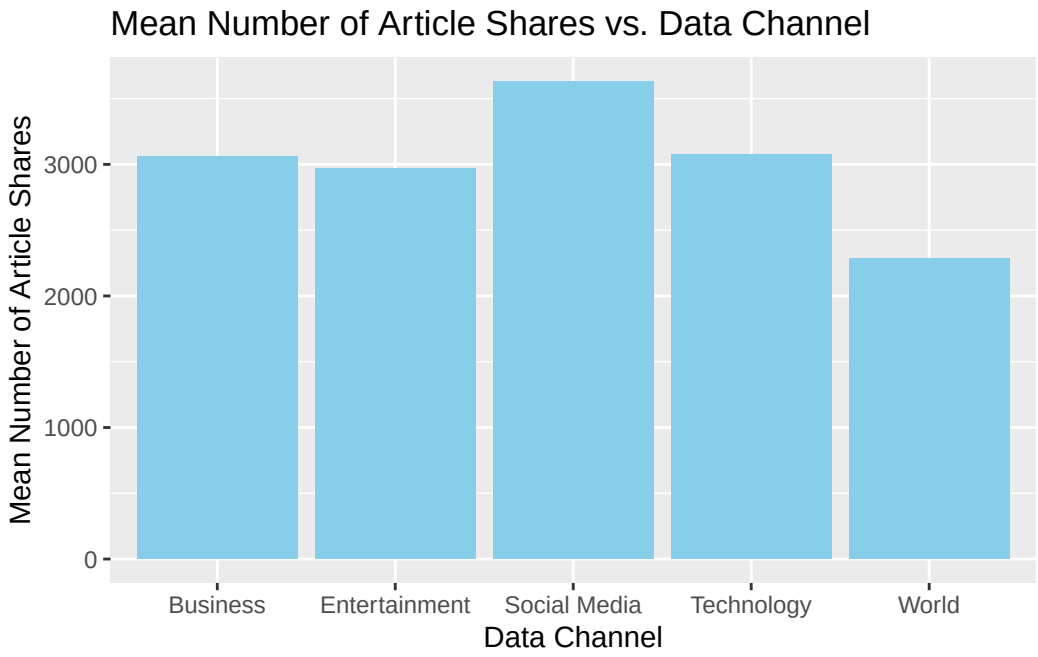
Bi-variable EDA

```
# A tibble: 6 x 2
  day_published mean_shares
  <fct>               <dbl>
1 Monday              3647.
2 Tuesday             3203.
3 Wednesday           3303.
4 Thursday            3179.
5 Friday              3285.
6 Saturday            4078.
```



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.

```
# A tibble: 6 x 2
  data_channel  mean_shares
  <fct>               <dbl>
1 Business            3063.
2 Entertainment       2970.
3 Social Media        3629.
4 Technology          3072.
5 World               2288.
6 <NA>                5368.
```




```
[1] 8233
```

We found that 8,233 articles are not tagged for a specific data channel, however, because we don't know if the cause of this missing tag (ex. the data being mistagged, etc.) we've decided to exclude the NA's from our data channel analysis altogether.

From our analysis, it appears that generally 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.