Value Types

Dataset Information

- The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.
- Acquisition date: January 8, 2015
  https://archive.ics.uci.edu/dataset/332/online+news+popularity
- The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method. See their article for more details on how the relative performance values were set.

Attribute Information:

```
 0.url: URL of the article (non-predictive)
 1. timedelta:                     Days between the article publication and the dataset
 2. n_tokens_title:                Number of words in the title
 3. n_tokens_content:              Number of words in the content
 4. n_unique_tokens:               Rate of unique words in the content
 5. n_non_stop_words:              Rate of non-stop words in the content
 6. n_non_stop_unique_tokens:      Rate of unique non-stop words in the content
 7. num_hrefs:                     Number of links
 8. num_self_hrefs:                Number of links to other articles published by Mashable
 9. num_imgs:                      Number of images
10. num_videos:                    Number of videos
11. average_token_length:          Average length of the words in the content
12. num_keywords:                  Number of keywords in the metadata
13. data_channel_is_lifestyle:     Is data channel ’Lifestyle’?
14. data_channel_is_entertainment: Is data channel ’Entertainment’?
15. data_channel_is_bus:           Is data channel ’Business’?
16. data_channel_is_socmed:        Is data channel ’Social Media’?
17. data_channel_is_tech:          Is data channel ’Tech’?
18. data_channel_is_world:         Is data channel ’World’?
19. kw_min_min:                    Worst keyword (min. shares)
20. kw_max_min:                    Worst keyword (max. shares)
21. kw_avg_min:                    Worst keyword (avg. shares)
22. kw_min_max:                    Best keyword (min. shares)
23. kw_max_max:                    Best keyword (max. shares)
24. kw_avg_max:                    Best keyword (avg. shares)
25. kw_min_avg:                    Avg. keyword (min. shares)
26. kw_max_avg:                    Avg. keyword (max. shares)
27. kw_avg_avg:                    Avg. keyword (avg. shares)
28. self_reference_min_shares:     Min. shares of referenced articles in Mashable
29. self_reference_max_shares:     Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess:    Avg. shares of referenced articles in Mashable
31. weekday_is_monday:             Was the article published on a Monday?
32. weekday_is_tuesday:            Was the article published on a Tuesday?
33. weekday_is_wednesday:          Was the article published on a Wednesday?
34. weekday_is_thursday:           Was the article published on a Thursday?
35. weekday_is_friday:             Was the article published on a Friday?
36. weekday_is_saturday:           Was the article published on a Saturday?
```

```
37. weekday_is_sunday:             Was the article published on a Sunday?
38. is_weekend:                    Was the article published on the weekend?
39. LDA_00:                        Closeness to LDA topic 0
40. LDA_01:                        Closeness to LDA topic 1
41. LDA_02:                        Closeness to LDA topic 2
42. LDA_03:                        Closeness to LDA topic 3
43. LDA_04:                        Closeness to LDA topic 4
44. global_subjectivity:           Text subjectivity
45. global_sentiment_polarity:     Text sentiment polarity
46. global_rate_positive_words:    Rate of positive words in the content
47. global_rate_negative_words:    Rate of negative words in the content
48. rate_positive_words:           Rate of positive words among non-neutral tokens
49. rate_negative_words:           Rate of negative words among non-neutral tokens
50. avg_positive_polarity:         Avg. polarity of positive words
51. min_positive_polarity:         Min. polarity of positive words
52. max_positive_polarity:         Max. polarity of positive words
53. avg_negative_polarity:         Avg. polarity of negative  words
54. min_negative_polarity:         Min. polarity of negative  words
55. max_negative_polarity:         Max. polarity of negative  words
56. title_subjectivity:            Title subjectivity
57. title_sentiment_polarity:      Title polarity
58. abs_title_subjectivity:        Absolute subjectivity level
59. abs_title_sentiment_polarity:  Absolute polarity level
60. shares:                        Number of shares (target)
```

Numeric:

timedelta, n_tokens_title, n_tokens_content, n_unique_tokens, n_non_stop_words, n_non_stop_unique_tokens, num_hrefs, num_self_hrefs, num_imgs, num_videos, average_token_length, num_keywords, global_subjectivity, global_sentiment_polarity, global_rate_positive_words, global_rate_negative_words, shares, etc.

Categorical:

url, data_channel_is_lifestyle, data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, data_channel_is_world, weekday_is_monday, weekday_is_tuesday, etc. Boolean: is_weekend

Coding Schemes

data_channel_is_lifestyle, data_channel_is_entertainment, etc. use 1 for “Yes” and 0 for “No”. weekday_is_monday, weekday_is_tuesday, etc. are using 1 to represent “True” and 0 to represent “False”.

Data Quantity

Format: The data is in CSV format. The dataset of Mashable articles written before 2015, obtained from the UCI Machine Learning Repository

Database Size

Dataset contains 61 columns and has 39,644 unique values.

Data Quality

The data includes multiple characteristics that could be relevant to the business question of what contributes to an article's popularity. Are certain topics more popular than others? What days of the week should we publish articles to get the most shares?

variables below might be useful: n_tokens_content, global_rate_negative_words, weekday_is_monday, weekday_is_tuesday, etc. is_weekend, num_shares, num_videos