0. url: URL of the article

1. timedelta: Days between the article publication and

the dataset acquisition---Target Leakage variable

2. n\_tokens\_title: Number of words in the title—can be considered in the model

3. n\_tokens\_content: Number of words in the content—can be considered in the model

4. n\_unique\_tokens: Rate of unique words in the content—can be considered in the model

5. n\_non\_stop\_words: Rate of non-stop words in the content-- can be considered in the model

6. n\_non\_stop\_unique\_tokens: Rate of unique non-stop words in the

Content-- can be considered in the model

7. num\_hrefs: Number of links-- can be considered in the model

8. num\_self\_hrefs: Number of links to other articles

published by Mashable-- can be considered in the model

9. num\_imgs: Number of images-- can be considered in the model

10. num\_videos: Number of videos-- can be considered in the model

11. average\_token\_length: Average length of the words in the

Content-- can be considered in the model

12. num\_keywords: Number of keywords in the metadata--?

13. data\_channel\_is\_lifestyle: Is data channel 'Lifestyle'?—can be considered

14. data\_channel\_is\_entertainment: Is data channel 'Entertainment'?-- can be considered in the model

15. data\_channel\_is\_bus: Is data channel 'Business'?-- can be considered in the model

16. data\_channel\_is\_socmed: Is data channel 'Social Media'?-- can be considered in the model

17. data\_channel\_is\_tech: Is data channel 'Tech'?-- can be considered in the model

18. data\_channel\_is\_world: Is data channel 'World'?-- can be considered in the model

19. kw\_min\_min: Worst keyword (min. shares)-- can be considered in the model

20. kw\_max\_min: Worst keyword (max. shares)-- can be considered in the model

21. kw\_avg\_mi

n: Worst keyword (avg. shares)-- can be considered in the model

22. kw\_min\_max: Best keyword (min. shares)-- can be considered in the model

23. kw\_max\_max: Best keyword (max. shares)-- can be considered in the model

24. kw\_avg\_max: Best keyword (avg. shares)-- can be considered in the model

25. kw\_min\_avg: Avg. keyword (min. shares)-- can be considered in the model

26. kw\_max\_avg: Avg. keyword (max. shares)-- can be considered in the model

27. kw\_avg\_avg: Avg. keyword (avg. shares)-- can be considered in the model

28. self\_reference\_min\_shares: Min. shares of referenced articles in

Mashable: It refers to minimum hits of similar articles shared.—can be considered in the model

29. self\_reference\_max\_shares: Max. shares of referenced articles in

Mashable—It refers to maximum hits of similar articles shared.—can be considered in the model

30. self\_reference\_avg\_sharess: Avg. shares of referenced articles in

Mashable-- It refers to average hits of similar articles shared.—can be considered in the model

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?

Please see below notes for the above variables’ explanation and inclusion in the model

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

Not sure what LDA means

44. global\_subjectivity: Text subjectivity-can consider in the model

45. global\_sentiment\_polarity: Text sentiment polarity- can consider in the model

46. global\_rate\_positive\_words: Rate of positive words in the content- There can be positive, negative or neutral words. This variable is total number of positive words in the content. can consider in the model

47. global\_rate\_negative\_words: Rate of negative words in the content- There can be positive, negative or neutral words. This variable is total number of negative words in the content. can consider in the model

48. rate\_positive\_words: Rate of positive words among non-neutral

Tokens- This explains positive words among non neutrals and the rate of positive and negative will be 1. can consider in the model

49. rate\_negative\_words: Rate of negative words among non-neutral

Tokens- This explains negative words among non neutrals and the rate of positive and negative will be 1. can consider in the model

50. avg\_positive\_polarity: Avg. polarity of positive words - can consider in the model. Average is average of minimum and max.

51. min\_positive\_polarity: Min. polarity of positive words - can consider in the model. Positive polarity range from 0 to 1. Minimum polarity refers to minimum positive value of words

52. max\_positive\_polarity: Max. polarity of positive words- can consider in the model. Maximum polarity refers to maximum positive value of words

53. avg\_negative\_polarity: Avg. polarity of negative words- can consider in the model. Average is average of minimum and max.

54. min\_negative\_polarity: Min. polarity of negative words. Negative polarity range from 0 to -1. Minimum polarity refers to minimum negative value of words. can consider in the model

55. max\_negative\_polarity: Max. polarity of negative words. Max negative value in range of 0 to -1. can consider in the model

56. title\_subjectivity: Title subjectivity. can consider in the model

57. title\_sentiment\_polarity: Title polarity. can consider in the model

58. abs\_title\_subjectivity: Absolute subjectivity level. can consider in the model

59. abs\_title\_sentiment\_polarity: Absolute polarity level. can consider in the model

60. shares: Number of shares (target)-

We can categorize weekdays and weekend into 1 category of weekend and non weekend. We will run model with all weekdays and only option of weekend and non weekend

Out of average, min and max, we can consider only average as avg is sum of min and max by 2 otherwise it’s multicollinearity.

Not sure about what is LDA and metadata.