# **Online News Popularity**

# **Final Project**

# **Sonya Tahir**

# Introduction

We have entered into the Information Age which has brought with it the Digital Revolution. People are becoming more and more comfortable with using computers and mobile phones in their everyday life for more than their traditional uses. They are being used to watch videos, listen to music, read books, do shopping, make financial transactions, plan and make travel arrangements, connect with friends and family, learn new skills, etc. This shift has come about due to accessibility of such technology to everybody as well as the convenience that they provide to people.

I want to focus on one particular use of the internet and the digital age and that is online news. Traditional news channels now maintain their websites and update the latest news online. Many other websites have also sprung up containing news related to a wide variety of areas and topics. The news websites have taken on the feel of a magazine that is always up-to-date and that can bring to the reader the news from his area of interest. “Past studies in this field have shown that there are 9 socio-technical advantages that have determined the adoption and use of online news: no costs, multitasking, more news choices, in-depth and background information, 24/7 updates, customization, ability to discuss the news with peers, the existence of different viewpoints, and the opportunity to ‘talk back to the media’.

Out of the 9 attributes, immediacy seemed to be the main reason for online news adoption. 70% of online news users had visited news sites a few times a day, while 47% of them would go to the internet first if they found out something interesting had happened.” (Online Journalism Blog, n.d.)

The availability of the latest news whenever the reader wishes is the most attractive reason for the adoption of online news. The reader can customize the topics of his interest to be shown related articles only. Websites also allow their articles to be shared on Facebook and Twitter due to which the reader can share articles with friends and family who may be interested in reading the same article. The sharing of articles results in other people reading the news even if they haven’t subscribed to the news website. More readers might visit the webpage due to the sharing of articles and this generates traffic to the website. The websites with high traffic are attractive to advertisers. All websites are interested in increasing webpage traffic as to attract advertisers and increase their revenue. For news websites, it is important to know which articles are the most popular so that they may sell ad space on the same page. They may even price it highly on an article that is expected to be popular. The number of page views and the number of page shares are important indicators of popularity of an article.

For the purpose of this project, I want to analyze what attributes of an article result in making it more popular. The research question is: Can we predict in advance which article will be more popular based on some characteristics of the article? In other words, what are the number of shares expected for each new article?

# Dataset

I will be looking at the data from Mashable which is “a leading global media company that informs, inspires and entertains the digital generation. It has an audience of 45 million monthly unique visitors and 27 million social followers.” (Mashable, n.d.) The data was collected on May 31, 2015 and summarizes a set of characteristics of approximately 40,000 articles published by Mashable in a period of two years. The data has 58 predictive attributes and 1 target field which is the number of times the article was shared on social media. (UCI, n.d.)

The description of each variable in the dataset is given in Appendix A. The target variable as well as all the independent variables are numeric. Some of the independent variables are binary.

# Exploratory Analysis

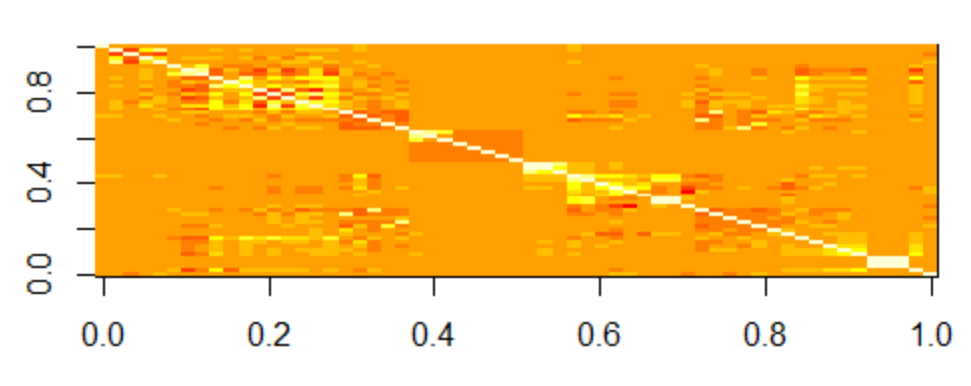
## Summary and Scatterplots

To get a fair idea about the variables in the dataset and their distributions, I observed the summary of the dataset.

Scaling of the variables?

I also tried to plot the relationships between the variables using scatterplots. However, since the number of variables is large, it is difficult to plot 58x58 combinations in scatterplots. Therefore, I preferred to look at the correlation figures instead of the scatterplots to see the relationship between the variables.

## Correlation and VIF Scores



The image shows that there are significant correlations in many variables. To find out which variables, I looked at the correlation matrix. I also calculated the VIF scores for the variables to decide which variables to remove from the dataset.

The correlation values show that n\_unique\_tokens is highly correlated with n\_non\_stop\_words and n\_non\_stop\_unique\_tokens. These are similar metrics anyway and we can use only n\_unique\_tokens in the final analysis and remove the other two variables. Similarly kw\_max\_min (max shares for worst keyword) and kw\_avg\_min (average shares for worst keyword) are highly correlated. kw\_min\_min (min shares for the worst keyword) is highly correlated with kw\_max\_max (max shares for the best keyword). The VIF scores also show values greater than 10 for some of these variables. By removing the variables for the average and keeping the min and max, I obtained VIF scores for the min and max variables less than 10. Similarly, the variables self\_reference\_min\_shares and self\_reference\_max\_shares are both highly correlated with self\_reference\_avg\_shares. I removed average and kept the min and max. Rate\_positive\_words and rate\_negative\_words also have high VIF. Similar variables of global\_rate\_positive\_words and global\_rate\_negative\_words are also present in the data. Therefore the variables with high VIF were removed. There are min, max and avg variables for negative polarity and positive polarity also. However, these are not highly correlated and do not have high VIF scores. Therefore I have kept these variables in the dataset.

The dataset contains 7 binary variables as flags for each day of the week. It also has one flag for weekend. Also I was getting infinite VIF scores for all the binary weekday variables. This is because each variable could be determined exactly given the other 6. Apart from the infinite VIF scores, the forward stepwise selection method was also giving error about linear dependence for these variables. I also converted the 7 binary variables into one categorical variable weekend. I kept the is\_weekend variable to experiment with later, in case each day information was not significant and just having weekend information was significant.

The dataset has 6 binary variables for data channel being lifestyle, entertainment, business, socialmedia, tech and world. I have also created a categorical variable data\_channel to represent this information in a more concise way.

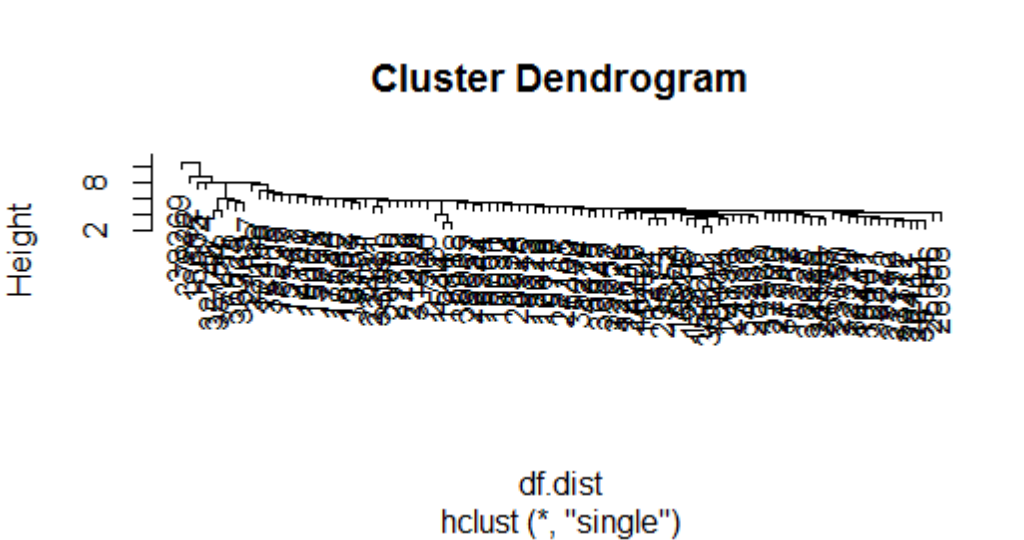
The dataset now has one target variable and 39 independent variables.

## PCA

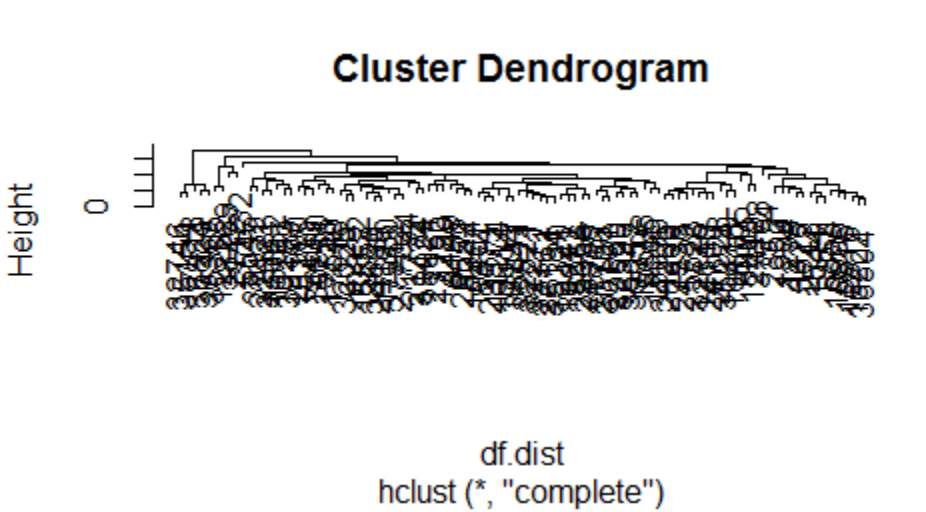
The first PC explains 11.65% variance of the data. The subsequent PCs explain lesser and lesser variance. These figures are not really significant. PCs would have been significant if just a few of them could have explained say 60% of the variance in the data. To explain approximately 60% variance in the data, we need to use the first 11 PCs. The dataset has 38 variables and using 11 PCs would be a way to reduce the number of variables. However, it appears that there is no significant underlying pattern in the data which can be captured using just one or two PCs. I therefore decided not to use PCs for prediction.

## Clustering

### Hierarchical Clustering



In the "single" clustering method, the observations seem to be evenly spread and do not seem to be forming distinct clusters. Initially it appears that three clusters are forming, but they get combined very soon and most of the other observations just keep getting added to the combined cluster.



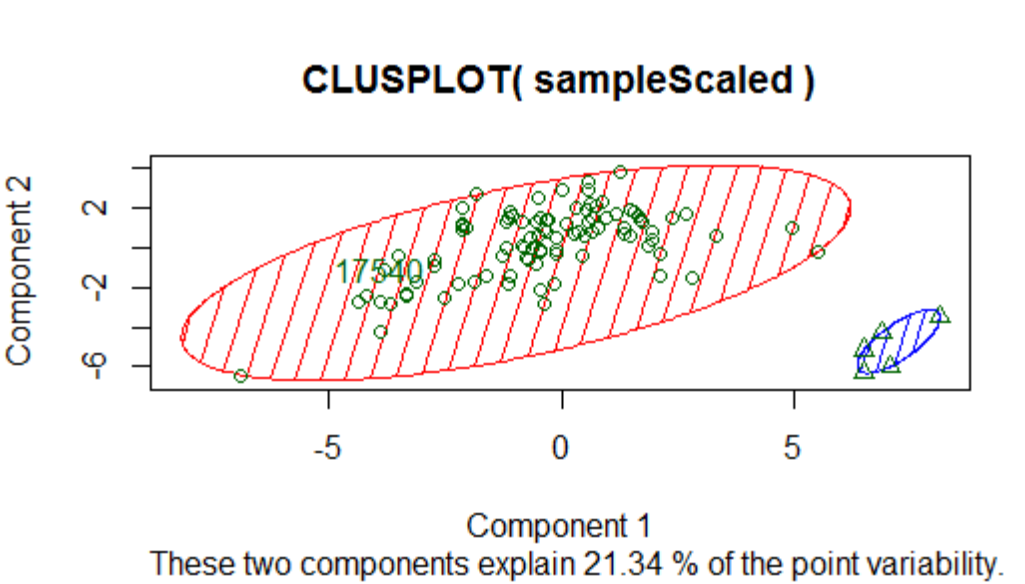
The "complete" clustering method gives a better picture. It seems that there is one large cluster and one small clusters in the data along with a few outliers. Most of the observations fall in one large cluster.

Taking the number of clusters as 5 gives us a large cluster, a small cluster, and few outliers. Thus it seems that most of the articles in the data are similar and get grouped together into one large cluster.

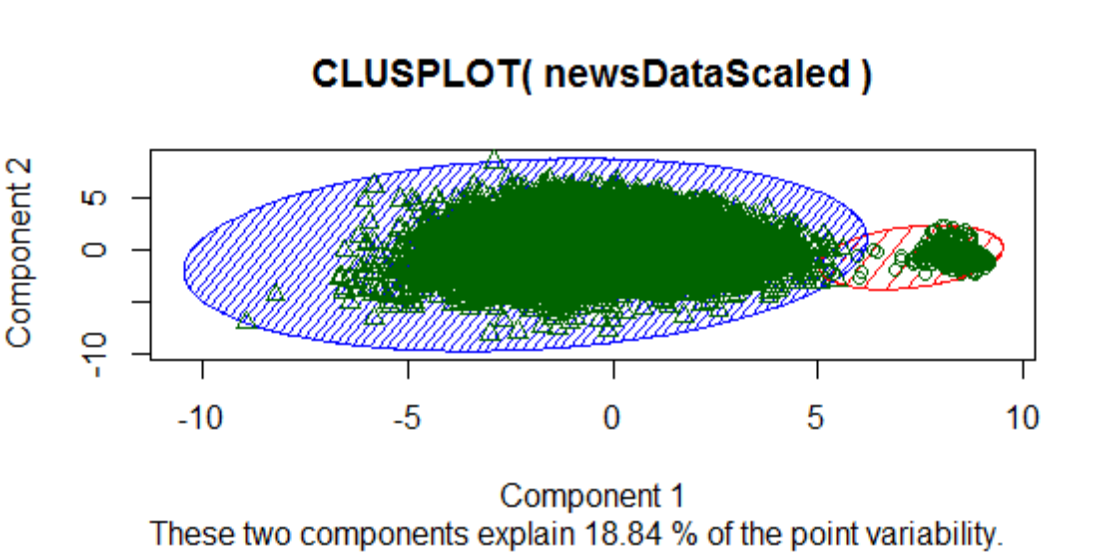
### K-Means Clustering

The within groups sum of squares keeps decreasing as we increase the number of clusters. Thus this confirms our previous observation that all the data points are scattered and they do not form clusters that are different from observations in other clusters.

Since we observed one large and one small cluster using hierarchical clustering (ignoring the outliers), we will consider k=2 as the best number of clusters and plot that on the first two PCs. The results show that there are two clusters in the data that are indeed quite different from each other. Most of the observations fall in the large cluster whereas few observations fall in the second cluster.



The above graph is for sample data of 100 rows only. However, the clusters for the full data are also similar with one large cluster and one small one. The clusters for the whole dataset are shown below:

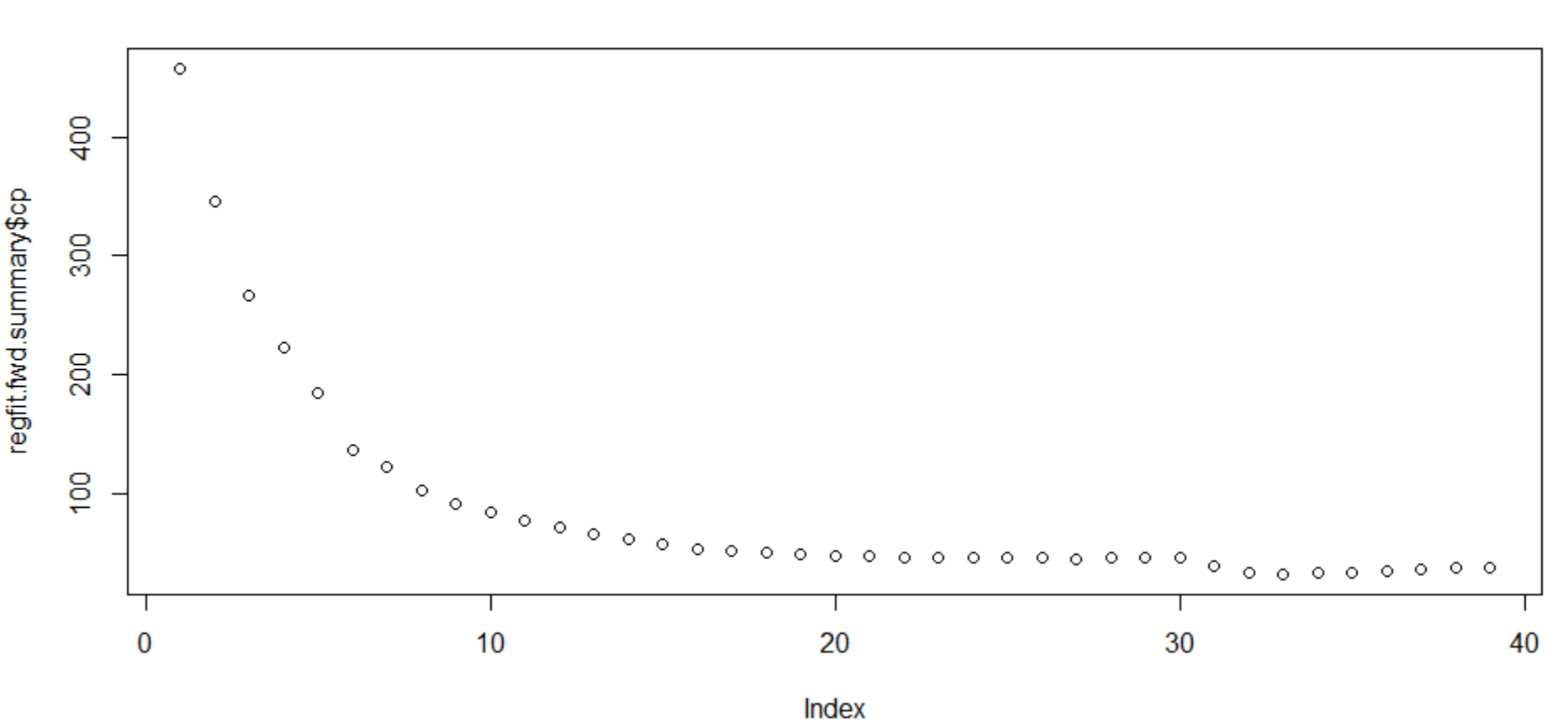


The proportion of rows in the two clusters is similar. The large cluster contains 38,452 rows and the small cluster contains 1192 rows. When we look at the observations in the small cluster, they have zeroes in many of the columns. Initially I thought that zero in many columns means that the data was not captured accurately and I considered deleting these rows. However, looking closely at the data in the small cluster, it appears that many of the columns have genuine values as well. The zeroes may therefore also be genuine values in case of these articles. 1192 is a large number of rows and the pattern of having zero in these columns may just be a quality that these documents share. Thus I decided not to delete these rows. I have also added the cluster ID as a variable to the dataset in case these two clusters have some predictive power for the number of shares. The dataset now has one target variable and 40 independent variables.

# Variable Selection

## Forward Stepwise Selection

The forward stepwise selection gives the following cp plot. The lowest cp is 32.32564 with 33 variables.

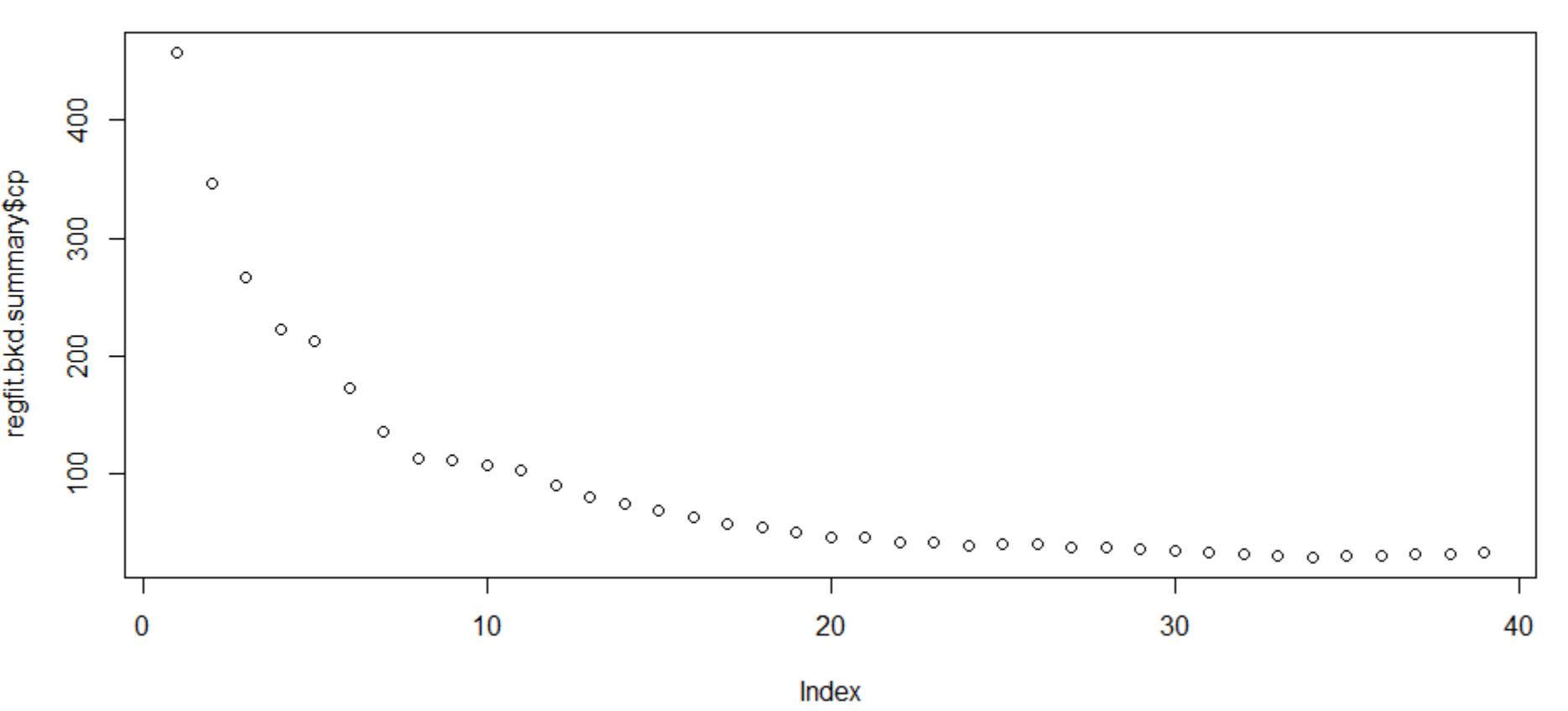


Since the weekday, data\_channel and cluster variables are categorical, they were specified as factors. Forward selection method treats each level of the categorical variable as a separate variable. If any level of the categorical level gets chose, I have considered it to be chosen into the model. The order of the variable selection is as follows:

LDA\_03, self\_reference\_min\_shares, kw\_max\_avg, num\_hrefs, LDA\_02, data\_channel, avg\_negative\_polarity, average\_token\_length, global\_subjectivity, num\_self\_hrefs, kw\_min\_avg, num\_imgs, n\_tokens\_title, num\_keywords, LDA\_01, self\_reference\_max\_shares, weekday, kw\_min\_max, min\_positive\_polarity, global\_rate\_positive\_words, abs\_title\_sentiment\_polarity, abs\_title\_subjectivity, kw\_min\_min,

## Backward Selection Method

The backward selection method gives the following cp plot. The lowest model has a cp of 29.29288 with 34 variables.



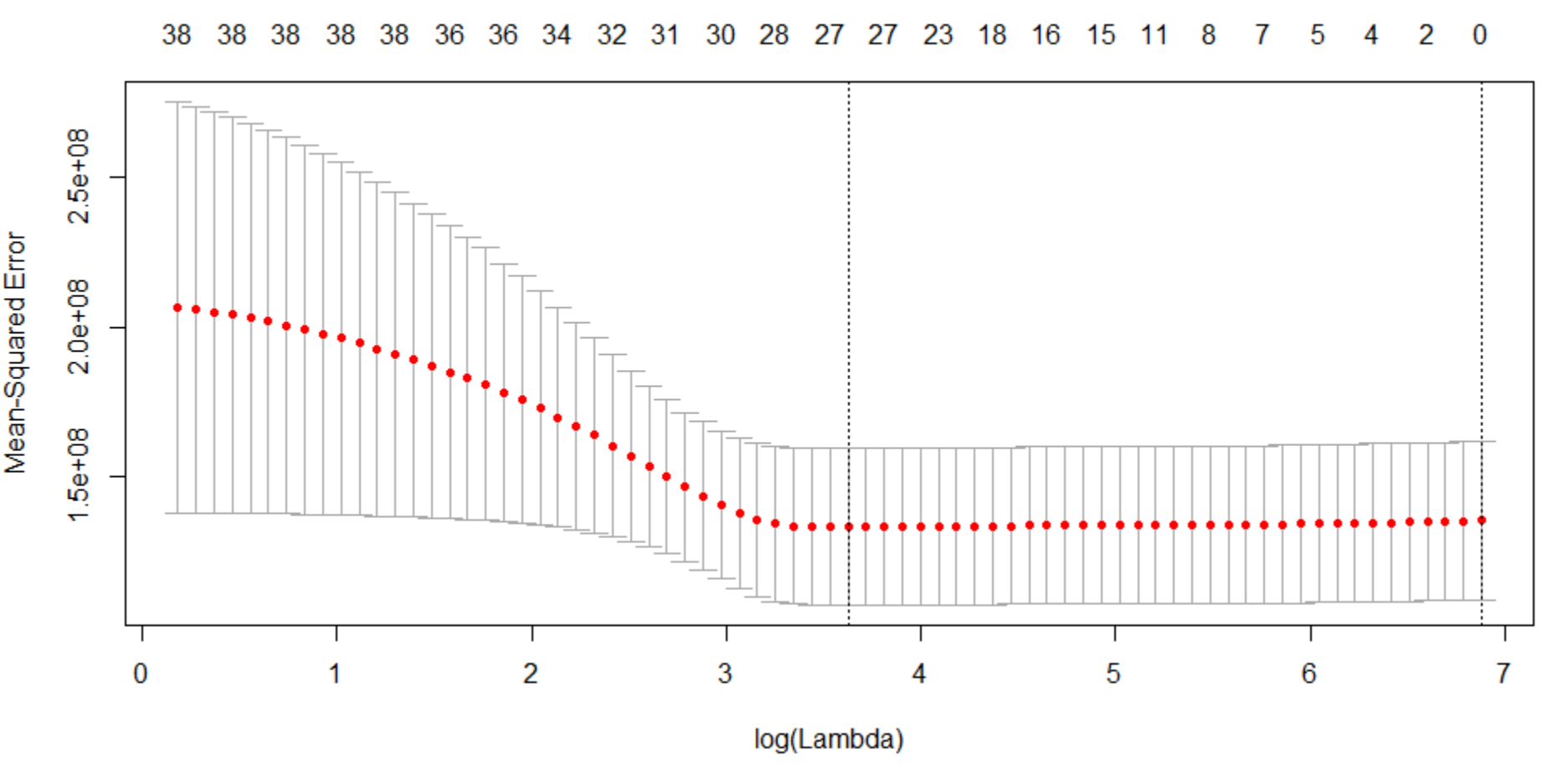
The variable selection is as follows:

LDA\_03, self\_reference\_min\_shares, kw\_max\_avg, num\_hrefs, average\_token\_length, global\_subjectivity, data\_channel, avg\_negative\_polarity, num\_self\_hrefs, kw\_min\_avg, num\_keywords, n\_tokens\_title, LDA\_00, num\_imgs, self\_reference\_max\_shares, LDA\_04, LDA\_01, min\_positive\_polarity, global\_rate\_positive\_words, n\_unique\_tokens, LDA\_02, n\_tokens\_content, weekday, abs\_title\_sentiment\_polarity, abs\_title\_subjectivity

The new variables selected by backward selection are LDA\_00, LDA\_04, n\_unique\_tokens, n\_tokens\_content.

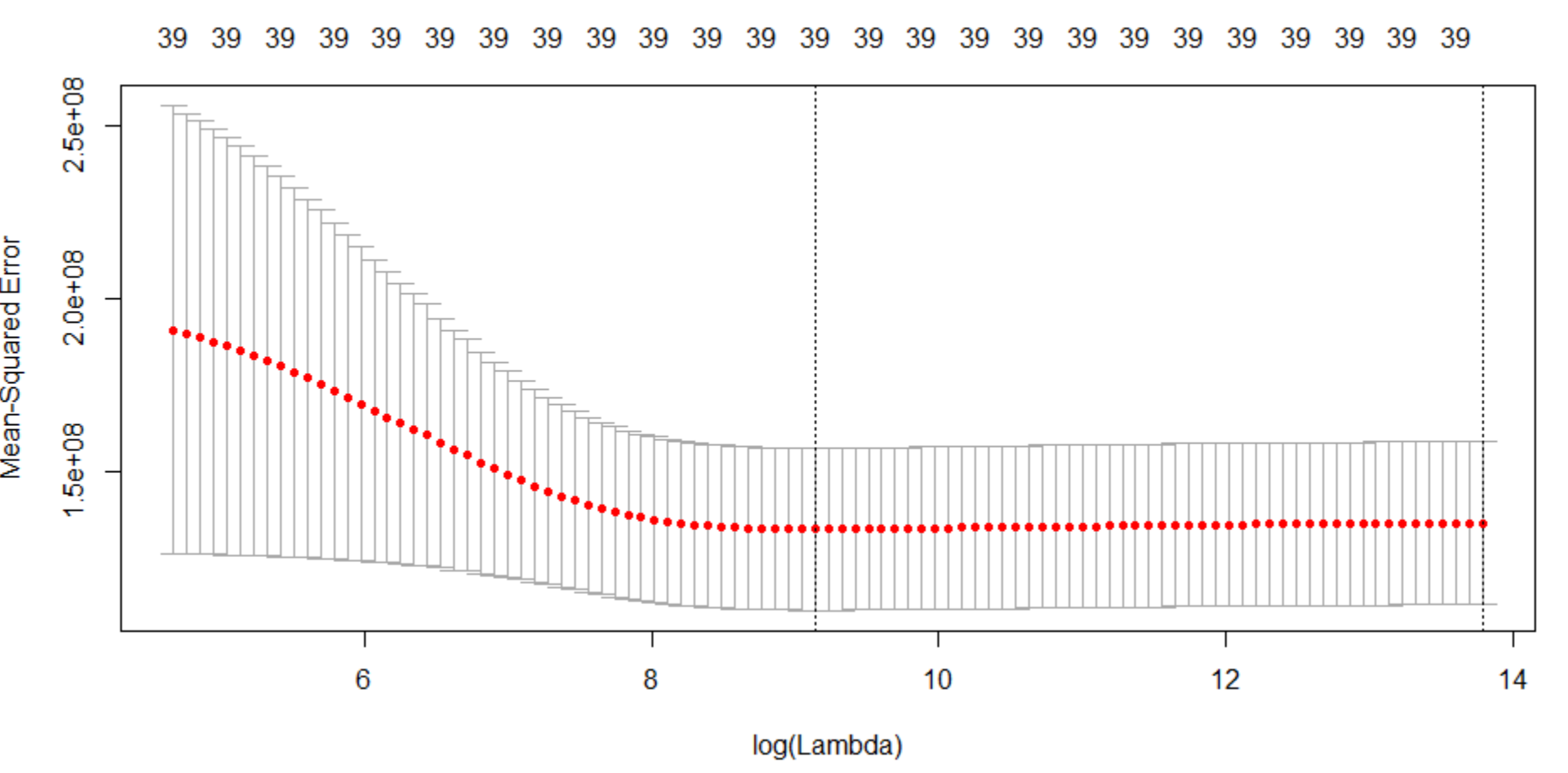
## LASSO

Lasso suggests using 27 variables.



## Ridge

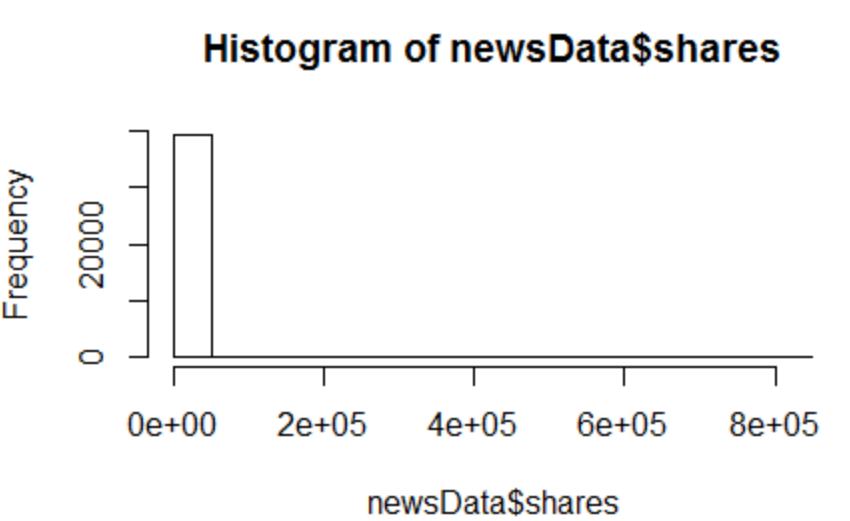
Ridge is not helpful as it selects all 39 variables.



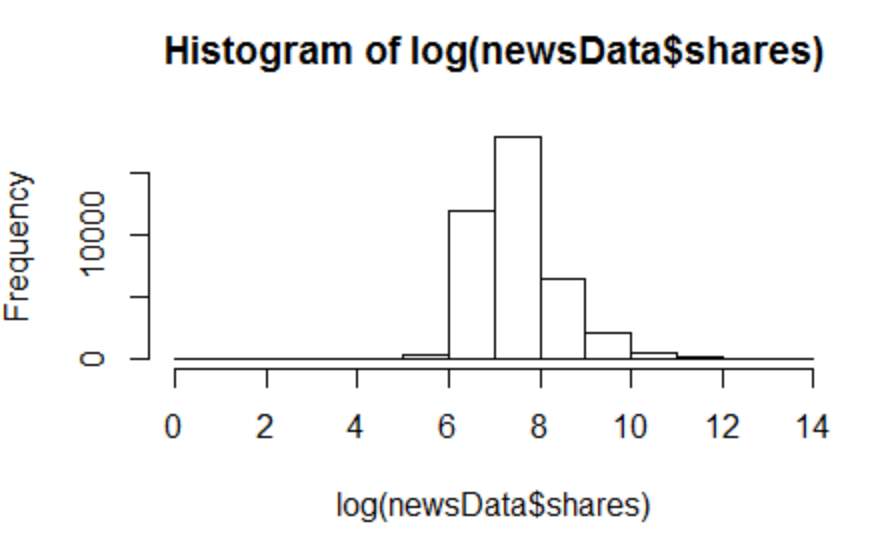
# Linear Regression

## Target Variable

The target variable “shares” is extremely skewed. There are some articles that have a very high number of shares. The distribution is given below:



Using the target as is was giving very high MSE in linear regression models (approximately 133480310). One option was to bin the data. However, knowing which articles get a very large number of shares is useful. I therefore used log(shares) as the target which was more evenly distributed. The distribution of log(shares) is given below:



## Variables from Forward and Backward Selection Methods

I tried linear regression model by adding each variable selected by forward selection method one by one and seeing if the mse improved. If the variable increased the mse, then it was not included in the model. Then I did the same with the variables selected by the backward selection model. In the third step, I combined the variables from both the methods and retained the ones that were giving me the lowest mse. I also deleted insignificant variables. The weekday variable that I created was significant for the weekdays and not for Saturday and Sunday. I therefore thought that this information could be captured by the is\_weekend variable and I replaced weekdays with is\_weekend. The is\_weekend variable was significant. The best MSE so far is 0.7824854.

## Polynomial Transformations

I applied polynomial transformations one by one on all numeric variables. Following are the results:

|  |  |
| --- | --- |
| LDA\_03  Degree Selected: 1 |  |
| self\_reference\_min\_shares  Degree Selected: 2 |  |
| kw\_max\_avg  Degree Selected: 6 |  |
| num\_hrefs  Degree Selected: 2 |  |
| LDA\_02  Degree Selected: 1 |  |
| avg\_negative\_polarity  Degree Selected: 1 |  |
| average\_token\_length  Degree Selected: 4  Variable was removed. |  |
| global\_subjectivity  Degree Selected: 3 |  |
| num\_self\_hrefs  Degree Selected: 2 |  |
| kw\_min\_avg  Degree Selected: 4 |  |
| min\_positive\_polarity  Degree Selected: 2 |  |
| global\_rate\_positive\_words  Degree Selected: 3  Final degree selected based on p-value: 2 |  |
| num\_keywords  Degree Selected: 4  Final degree selected based on p-value: 3 |  |
| num\_imgs  Degree Selected: 3  Final degree selected based on p-value: 3 |  |

The degree suggested by the polynomial transformation with the lowest MSE was applied on all the above variables. After this, the resultant model was observed and degrees were adjusted based on which ones were insignificant. The changes are also mentioned in the table above. The MSE obtained for this model was 0.764844.

## Splines

Using splines on all of the selected variables gives an MSE of 0.7538 which is better than the one obtained with polynomial transformations. I also tried using a combination of splines and polynomial transformations. The result was not better than the model using splines alone.

## Outliers

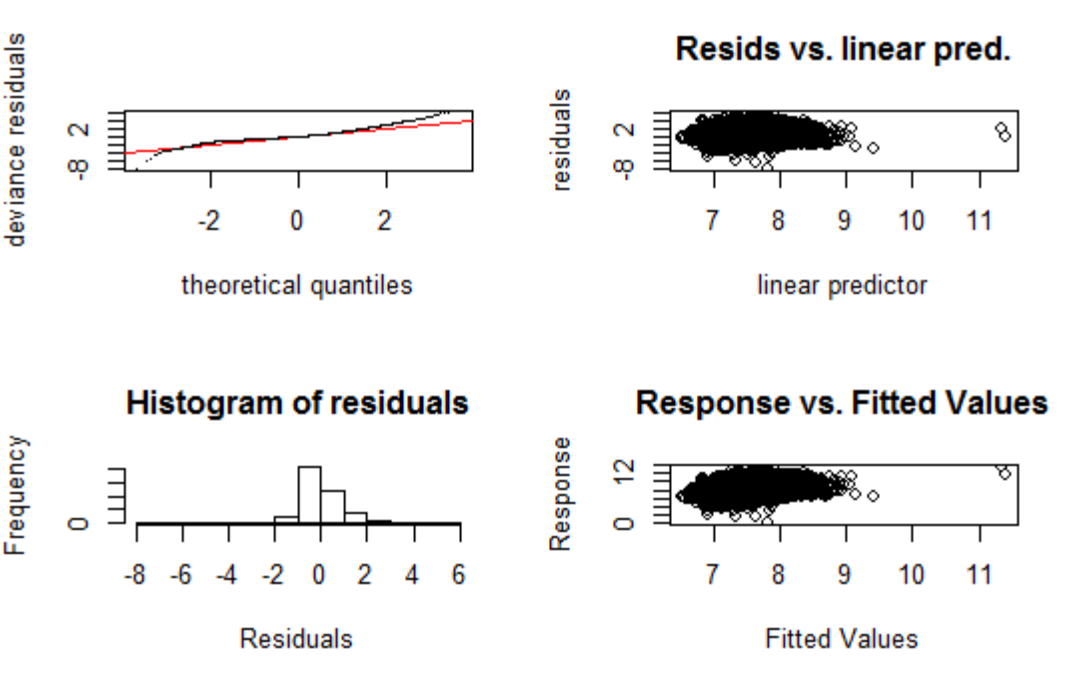
Outliers were identified and removed from the data. This decreased the MSE greatly. For lm5 with polynomial transformations, the MSE came down to 0.7580219. For lm6 with splines, the reduced MSE was 0.7525.

# Quantile Regression

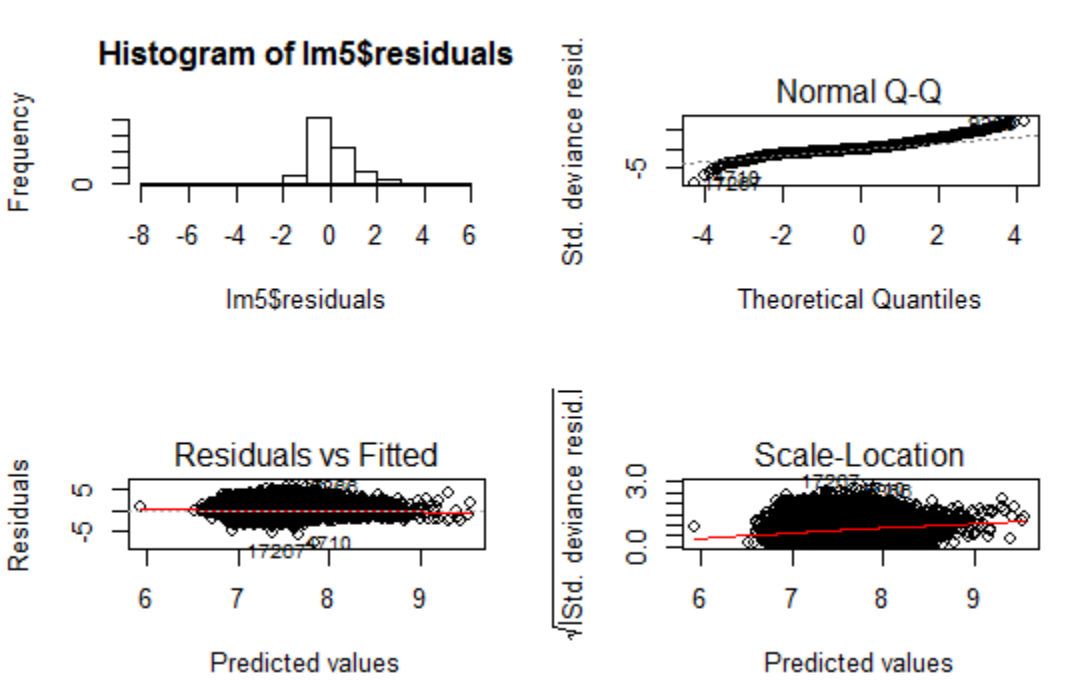
I also fitted quantile regression on the data using the quantreg library. I started from all the variables selected using the forward selection method and refined the model to get the lowest MSE. The MSE obtained was 0.8190282. Thus the best model is the one using splines.

# Model Assumptions

The best model is lm6 with splines. The diagnostics for this model are given below. The Shapiro Wilk’s test confirms that the residuals are not normal. The residuals are also not homoscedastic.



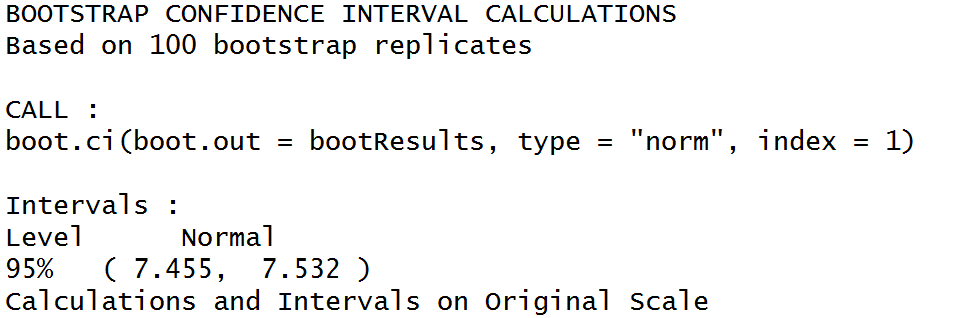
We can check the same for lm5 (model with polynomial transformations) which gave results similar to lm5. The diagnostic results are the same. The residuals are not normal as confirmed by the Shapiro Wilk’s test. The variance of the residuals is also not constant.



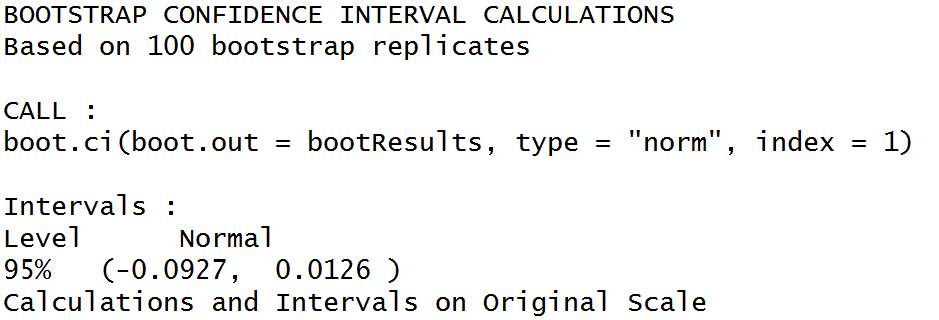
Thus it seems that linear regression is not the best suited model for this dataset. To model the target more closely, we may have to bin the target and use logistic regression. We can also use neural networks to predict the numeric target more accurately.

# Bootstrapping Confidence Intervals

Bootstrapping results for lm6 with splines is given below.



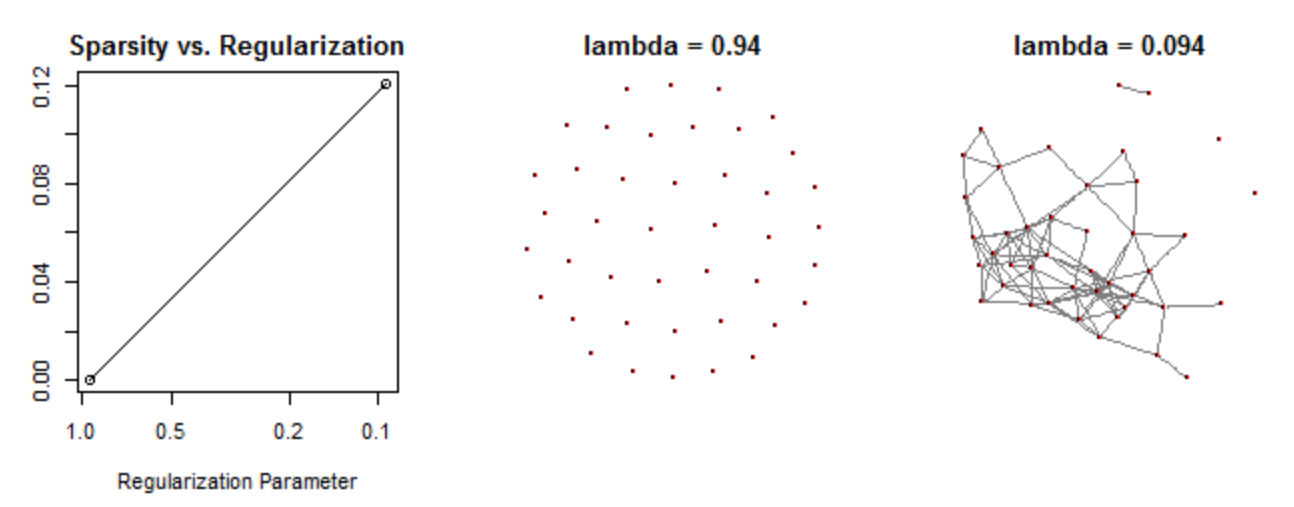
Bootstrapping results for lm5 with polynomial transformations is given below.



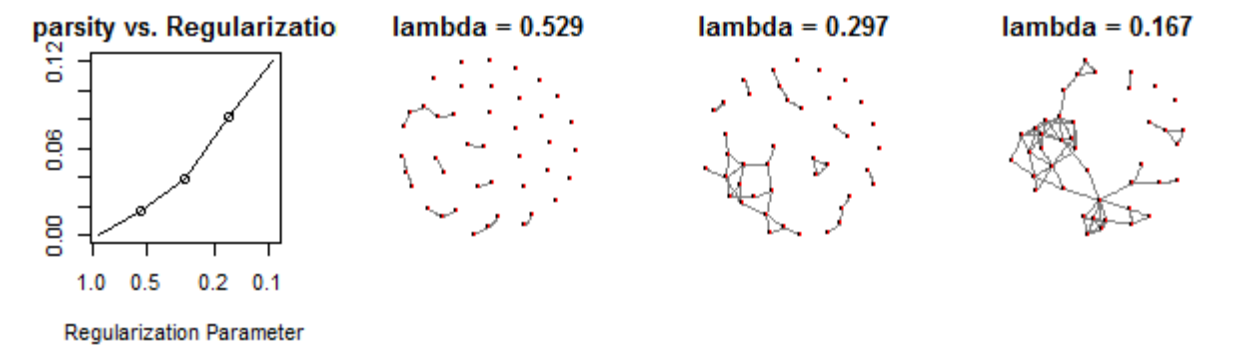
# Huge Package

Following graphs were created using Huge package.

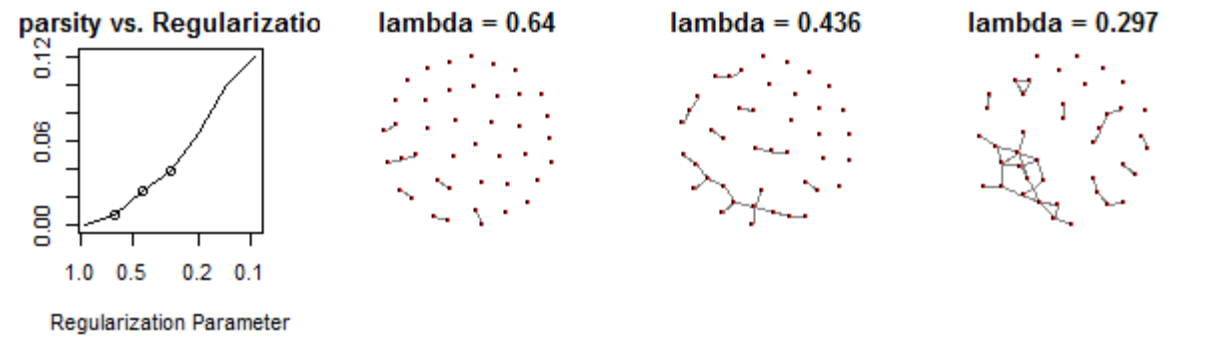
## Lambda = 2



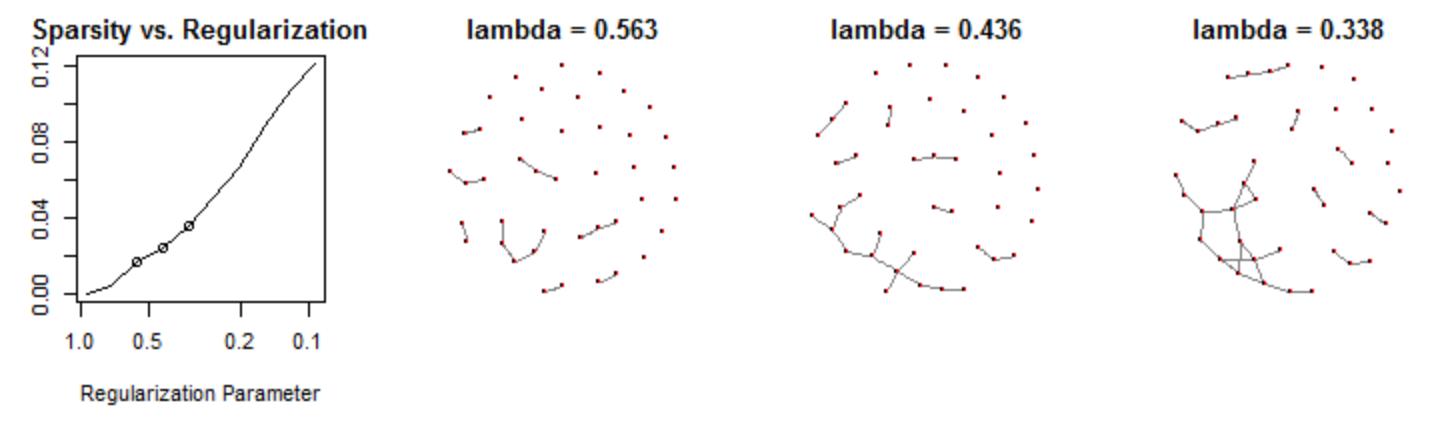
## Lambda = 5



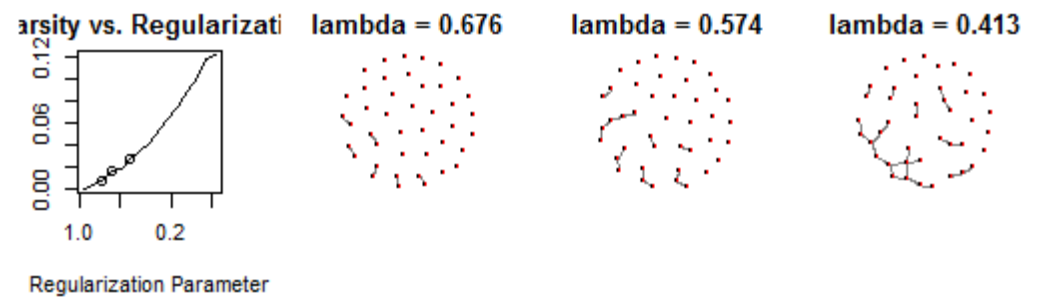
## Lambda = 7



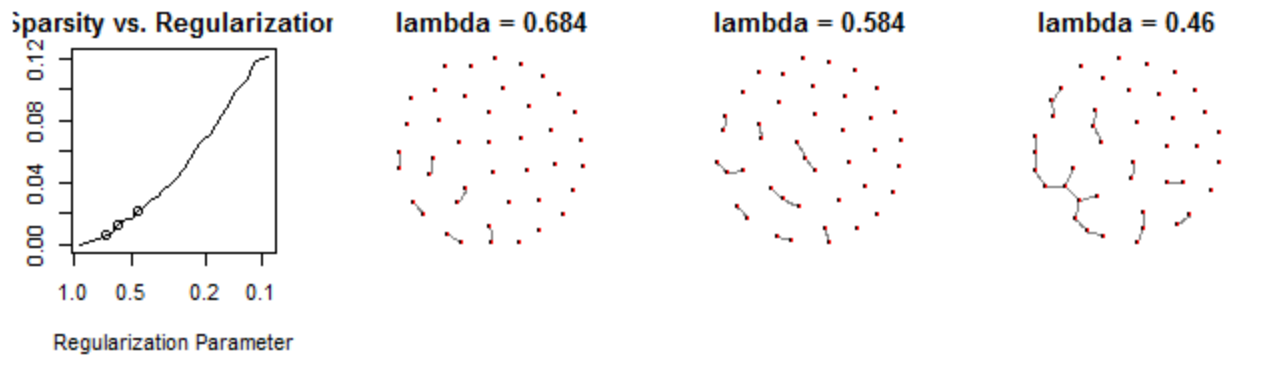
## Lambda = 10



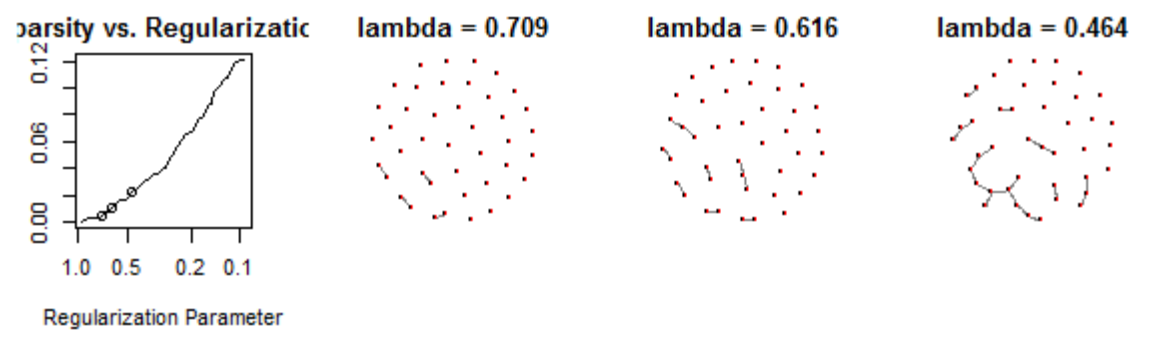
## Lambda = 15



## Lambda = 30



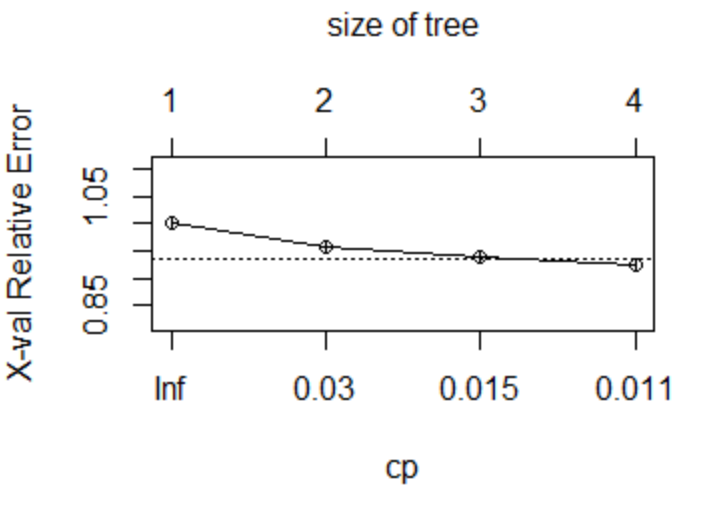
## Lambda = 50



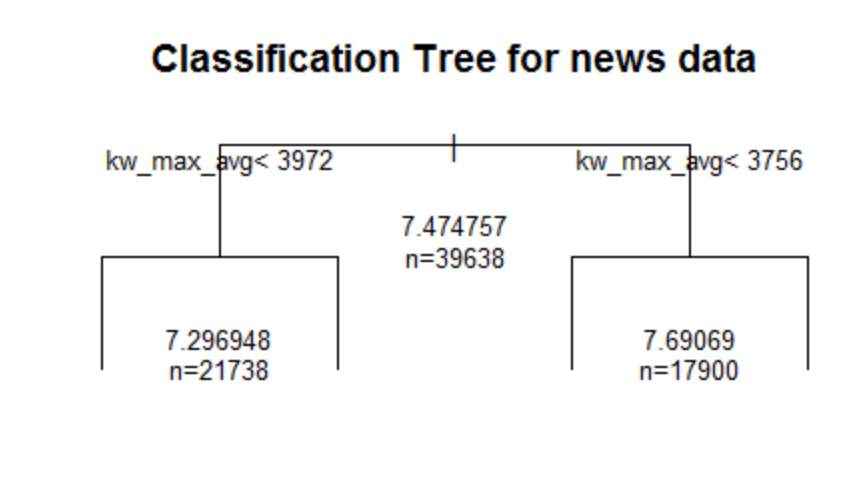
The results show that the data points are not connected to others in any pattern. There are no distinct groups having particular characteristics. Most of the data is dissimilar and disconnected. The results are similar to what I obtained using clustering. Most of the data points formed just one cluster which shows that there are no distinct groups in the data.

# Bonus Packages # 1, 2 rpart & rms

I used the rpart package to fit a regression tree on the data. The iterations with the corresponding cp are shown in the plot below:



The plot of the tree is given below:



To calculate the MSE, I used the rms library. The MSE for this model is 0.8426867408. This is worse than the MSE of the linear model. I will try random forests next.

# Bonus Package # 3 randomForest

I used the randomForest package to fit a random forest on my data. The mse for the fitted model is 0.7310637. The best model with splines had an MSE of 0.7525. The random forest has improved the MSE and is a better model.

# Bonus Package # 4 neuralnet

I used the neuralnet package to fit a neural network on my data. I used 1 hidden layer with 3 nodes. The MSE for this model was 161.1823238. Neural network can definitely perform better than this. This needs multiple trial and error to see how many hidden layers and how many layers in each node will optimize the model for this data.

# Conclusion

Out of all the models fitted on the data, the best model is the random forest with an MSE of 0.7310637. Out of the linear regression models, the best model was the one with splines which had an MSE of 0.7525. The linear model with polynomial transformations was a close third with an MSE of 0.7580219. Quantile regression, decision tree and neural network did not give good results. Neural network has the potential to give better results but requires more time for optimization.

# Appendix: Description of variables in the dataset

0. url: URL of the article (non-predictive)

1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)

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)

# References

“Why do people read online news?” Online Journalism Blog <http://onlinejournalismblog.com/2010/04/27/why-do-people-read-online-news-research-summary/> April 27, 2010

Mashable <http://mashable.com/about/>

# Dataset Link

“Online News Popularity Dataset” UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

# R Code

The R code for all the analysis has been submitted in the file FinalProject\_SonyaTahir.Rmd. It has not been added here as appendix because the code is more readable in an R file.