STAC67 Final Project

Alex Cheng, Zaamin Rattansi, Jacob Temple, Jeffrey Wong

12/05/2022

Introduction

The purpose of this research is to study the direct and indirect relationships of metrics for Mashable news articles and the amount an article was shared by readers. This will be done through thorough analysis of the variables that have an effect on the number of shares of an article. Variables that are directly related, and variables that have less importance were included to see the largest range of relationships possible and to really understand what affects the amount an article is shared by Mashable users.

Background

The data we have analyzed and studied is taken from a multi-platform media company called Mashable. The data was taken to analyze the number of shares a new article receives in relation to various other variables. The chosen variables had some relation to the amount an article may be shared and our models show the importance of some factors vs others regarding the amount an article may be shared. There may be many other factors that cause a post to be shared, however, we can only analyse the data that has been collected.

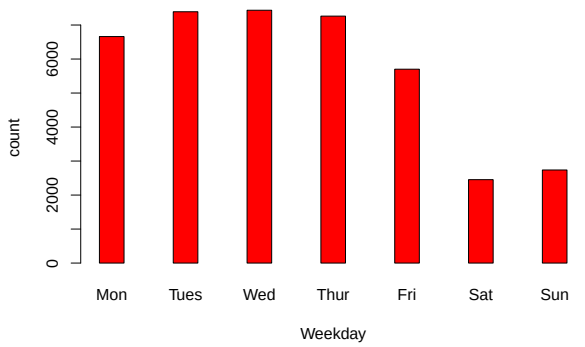
Study Goal

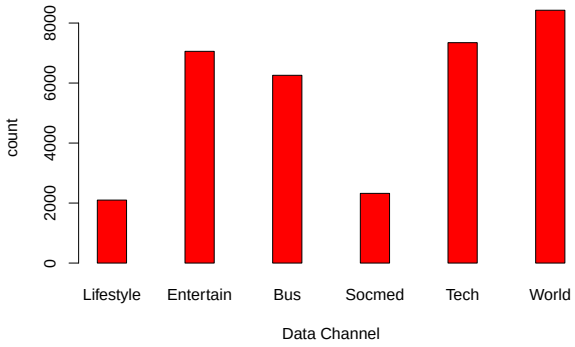
We will analyze what variables cause an increase or decrease in the number of shares to learn more about what future articles require in order to be shared more by Mashable users.

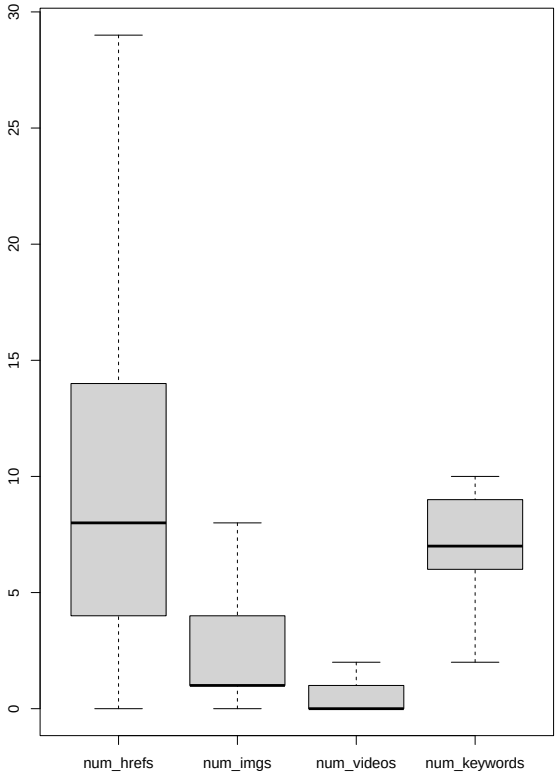
Description of Dataset

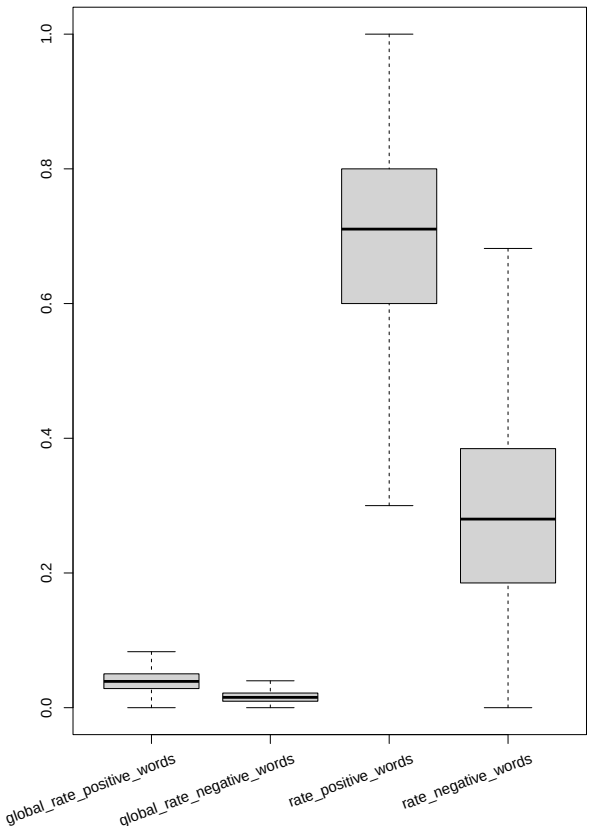
The data was cleaned mainly be eliminating variables that were either redundant to the number of shares or if similar variables were used. For example, we removed any max and min variables and instead used the average variables. It was unnecessary to use all three, the max, min, and average, when we can just use the average variables. We also removed all the LDA variables since they are irrelevant to the following analysis. It was also important to remove the baseline variables for categorical variables as the data already provides factored categorical variables (Weekday and Data Channel). Also removed were variables that had perfect collegiality, e.g., is_weekend.

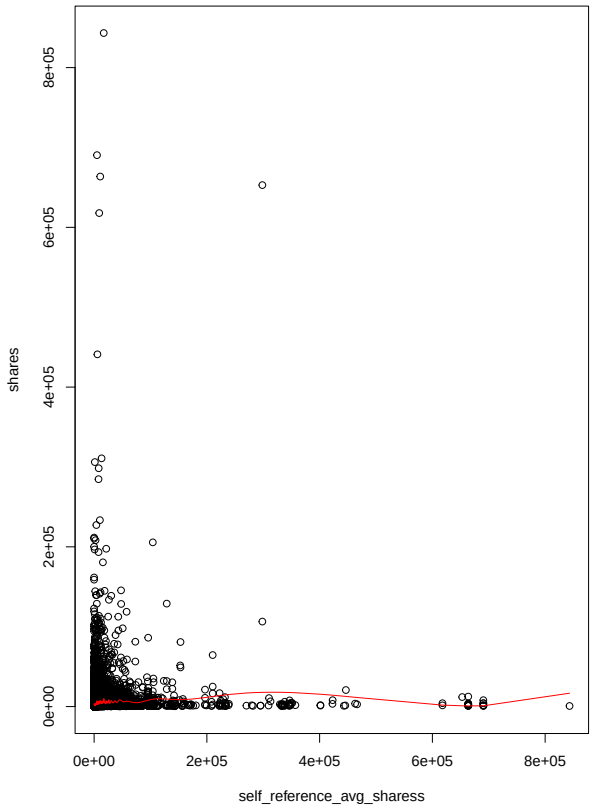
Preliminary Investigation of Data

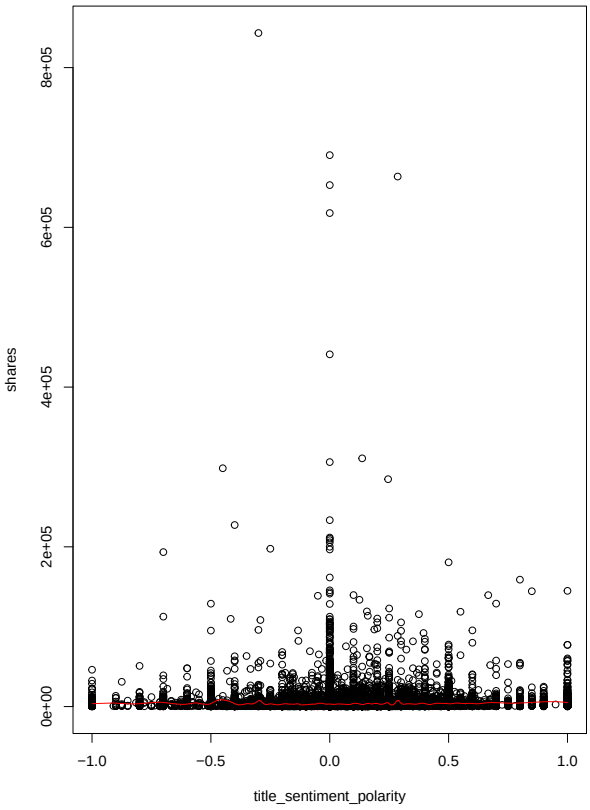


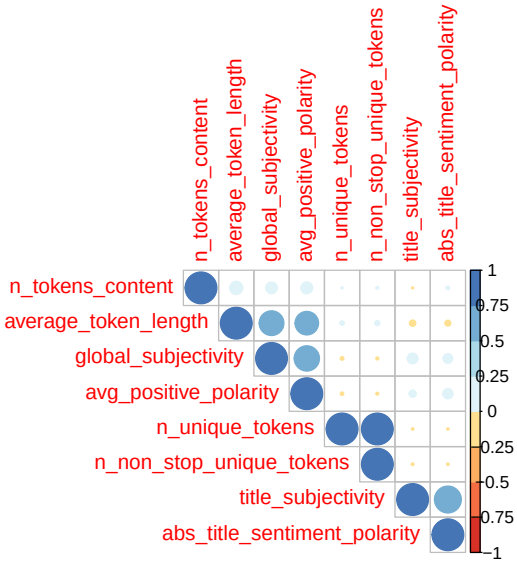







shares
0e+00
2e+05
4e+05
6e+05
8e+05
self_reference_avg_sharess


shares
0e+00
2e+05
4e+05
6e+05
8e+05
−1.0
−0.5
0.0
0.5
1.0
title_sentiment_polarity


n_tokens_content
average_token_length
global_subjectivity
avg_positive_polarity
n_unique_tokens
n_non_stop_unique_tokens
title_subjectivity
abs_title_sentiment_polarity
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1

Model Building

Full Model With Interactions

```
## $sigma
## [1] 10464.04
##
## $df
## [1]   125 19724   125
##
## $r.squared
## [1] 0.07146069
##
## $adj.r.squared
## [1] 0.06562319
##
## $fstatistic
##       value       numdf       dendf
##    12.24166   124.00000 19724.00000
```

Model 1 - Stepwise Regression of Full Model

```
## $sigma
## [1] 10533.72
##
## $df
## [1]    36 19813    36
##
## $r.squared
## [1] 0.05480761
##
## $adj.r.squared
## [1] 0.05313791
##
## $fstatistic
##       value       numdf       dendf
##    32.82486    35.00000 19813.00000
```

Model 2 - Subset Regression on Model 1

```
## $sigma
## [1] 10533.32
##
## $df
## [1]    30 19819    30
##
## $r.squared
## [1] 0.05459331
##
## $adj.r.squared
## [1] 0.05320995
##
```

```
## $fstatistic
##       value       numdf       dendf
##    39.4643    29.0000 19819.0000
```

Model Diagnostics

The number of outliers:

```
## [1] 96
```

Number of leverage points:

```
## [1] 915
```

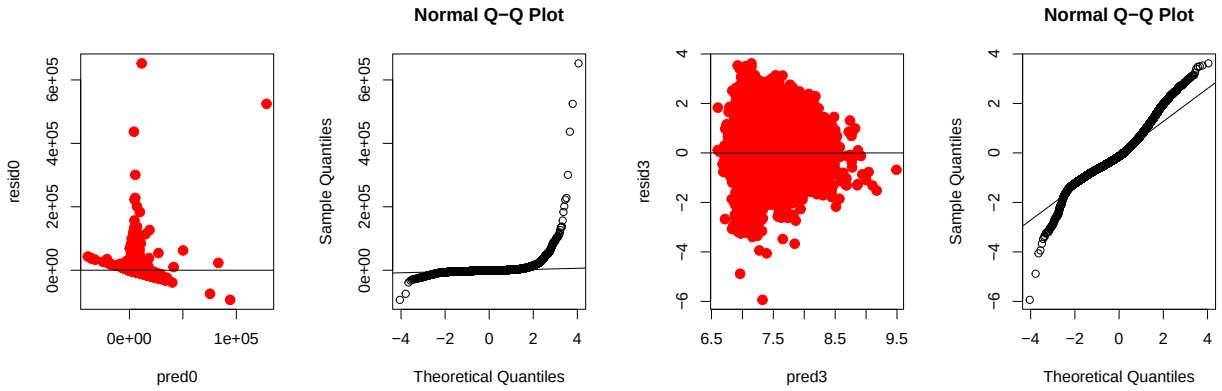
Number of influential points (DFFITS & DFBETAS):

```
## [1] 385
```

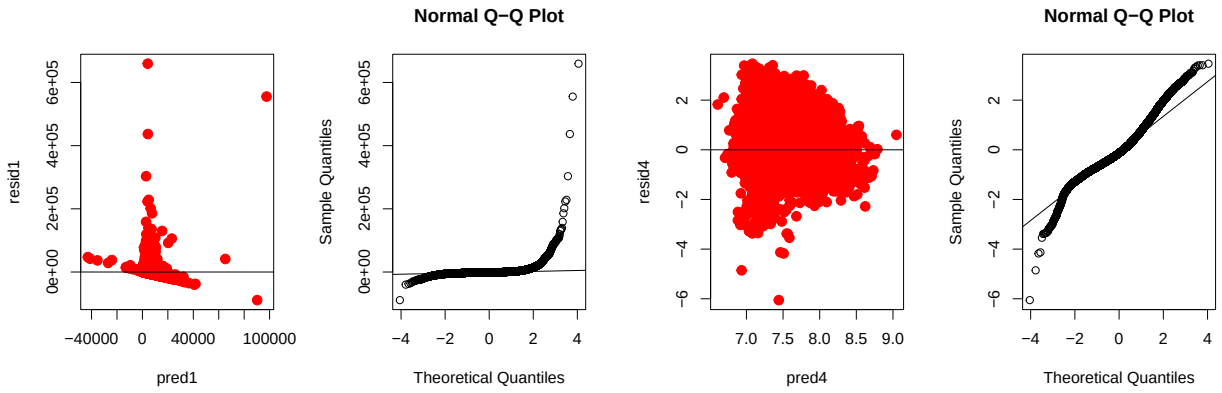
```
## [1] 7369
```

We removed all outliers, leverage points, and influential points.

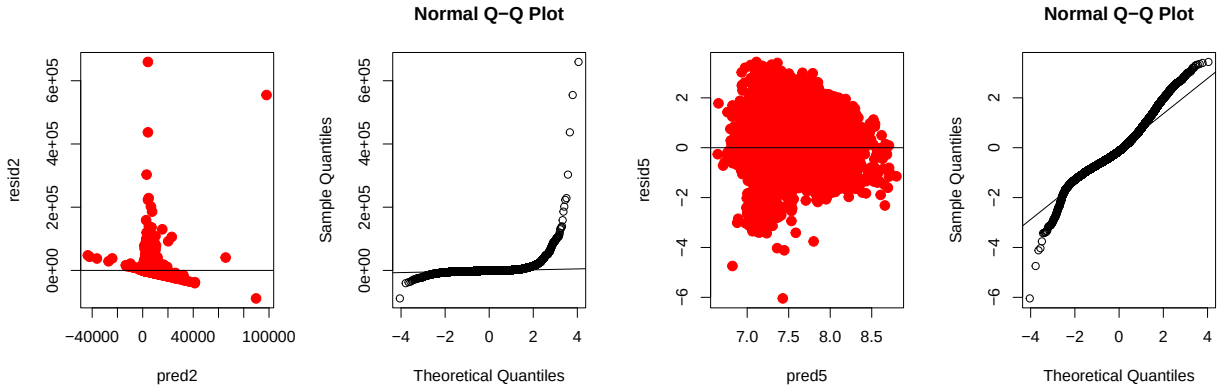
Model 3 - Full Model Before and After log Transformation



Model 4 - Model 1 Before and After log Transformation



Model 5 - Model 2 Before and After log Transformation



PRESS Statistics for Models 1-5

```
## [1] 4.24814e+13
```

```
## [1] 2.345082e+12
```

```
## [1] 2.340874e+12
```

```
## [1] 11593.87
```

```
## [1] 11993.85
```

```
## [1] 12073.8
```

Variables with VIF >10 for Model 5:

```
##                                                    [,1]
## self_reference_min_shares                             2
## self_reference_max_shares                             9
## self_reference_avg_sharess                           10
## self_reference_max_shares:global_subjectivity        19
## self_reference_max_shares:self_reference_avg_sharess 20
## self_reference_avg_sharess:global_subjectivity       22
## kw_avg_avg:self_reference_avg_sharess                23
## self_reference_avg_sharess:num_keywords              28
```

Model 6 - Model 5 After removing Variables with High VIF

```
## $sigma
## [1] 0.8053047
##
## $df
## [1]    22 18669    22
##
## $r.squared
## [1] 0.09032009
##
## $adj.r.squared
## [1] 0.08929683
##
## $fstatistic
##       value       numdf       dendf
##    88.26683    21.00000 18669.00000
```

Model Validation

Difference betwee MSPE and MSE for Models 3-6

```
## [1] 164935068
```

```
## [1] 164935151
```

```
## [1] 164935334
```

```
## [1] 164932973
```

Conclusion

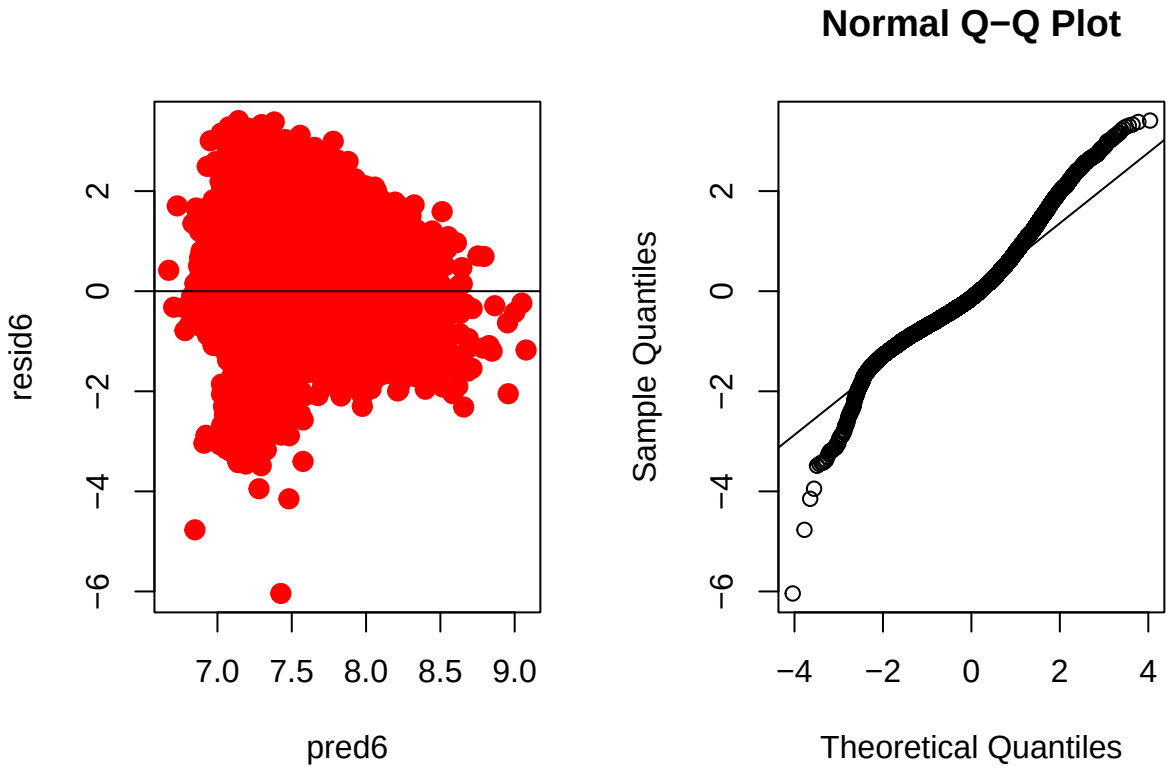
Final Model

Based on the above analysis, we have chosen the following linear model:

```
##
## Call:
## lm(formula = paste("log(shares) ~", vars6, sep = " "), data = new_build)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0402 -0.5270 -0.1366  0.4237  3.4092
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       7.010e+00  4.359e-02
## kw_avg_avg                                        2.118e-04  7.552e-06
## num_hrefs                                         5.300e-03  1.231e-03
## avg_negative_polarity                            -2.166e-01  5.666e-02
## average_token_length                             -2.657e-02  8.583e-03
## num_self_hrefs                                   -5.561e-03  4.348e-03
## weekday_is_monday                                -1.070e-01  4.114e-02
## data_channel_is_lifestyle                        -4.626e-02  2.707e-02
## kw_avg_max                                       -9.103e-07  6.257e-08
## num_hrefs:num_imgs                               -9.741e-05  8.022e-05
## average_token_length:num_imgs                     1.132e-03  3.299e-04
## avg_negative_polarity:data_channel_is_entertainment  6.829e-01  5.165e-02
## num_hrefs:num_self_hrefs                          1.587e-04  1.958e-04
## avg_negative_polarity:weekday_is_monday          -1.834e-01  1.441e-01
## average_token_length:kw_avg_min                  -7.964e-06  4.747e-06
## weekday_is_monday:self_reference_max_shares      -9.365e-06  3.794e-06
## weekday_is_monday:self_reference_avg_sharess      2.439e-05  6.132e-06
## self_reference_avg_sharess:self_reference_min_shares -1.766e-10  4.576e-11
## num_hrefs:self_reference_avg_sharess             -1.657e-08  1.157e-07
## num_self_hrefs:self_reference_avg_sharess         1.901e-06  4.851e-07
## kw_avg_max:self_reference_avg_sharess             3.438e-11  6.072e-12
## self_reference_avg_sharess:num_videos            -1.323e-07  2.082e-07
##                                                  t value Pr(>|t|)
## (Intercept)                                      160.829  < 2e-16 ***
## kw_avg_avg                                        28.038  < 2e-16 ***
## num_hrefs                                          4.304 1.68e-05 ***
## avg_negative_polarity                             -3.822 0.000133 ***
## average_token_length                              -3.096 0.001963 **
## num_self_hrefs                                    -1.279 0.200877
## weekday_is_monday                                 -2.602 0.009279 **
## data_channel_is_lifestyle                         -1.708 0.087563 .
## kw_avg_max                                       -14.548  < 2e-16 ***
## num_hrefs:num_imgs                                -1.214 0.224642
## average_token_length:num_imgs                      3.430 0.000605 ***
## avg_negative_polarity:data_channel_is_entertainment  13.222  < 2e-16 ***
## num_hrefs:num_self_hrefs                           0.811 0.417474
## avg_negative_polarity:weekday_is_monday           -1.273 0.203186
## average_token_length:kw_avg_min                   -1.678 0.093390 .
## weekday_is_monday:self_reference_max_shares       -2.469 0.013572 *
## weekday_is_monday:self_reference_avg_sharess       3.977 7.01e-05 ***
## self_reference_avg_sharess:self_reference_min_shares -3.860 0.000114 ***
## num_hrefs:self_reference_avg_sharess              -0.143 0.886090
## num_self_hrefs:self_reference_avg_sharess          3.920 8.89e-05 ***
## kw_avg_max:self_reference_avg_sharess              5.662 1.51e-08 ***
## self_reference_avg_sharess:num_videos             -0.635 0.525162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.8053 on 18669 degrees of freedom
## Multiple R-squared:  0.09032,    Adjusted R-squared:  0.0893
## F-statistic: 88.27 on 21 and 18669 DF,  p-value: < 2.2e-16
```



We can see that each variable only affects the number of share very slightly. Among the variables that affect the greatest change in shares, we see that with each unit increase in the average negative polarity of an article, while holding all other variables constant, we see a -0.22 decrease in the number of shares. Additionally, we see that articles posted on Mondays have 0.1 fewer shares than those posted on Sunday (baseline variable for the weekday on which an article was posted). We also see that the average negative polarity of articles posted in the entertainment data channel impacts the rate at which the article is shared. This aligns with our intuition of how an article's shares might be influenced.

Note that $R^2 = 0.09032$ and $R^2_{adj} = 0.0893$, which shows that the predictive power of our model is very low. Note also the difference between MSPE and MSE is very large.

As such, we have not found a model suitable for establishing a clear relationship between the variables captured in the data and the number of shares an article receives.

Obstacles

Notably missing are Brown-Forsyth and Shapiro-Wilks tests for equal and normally distributed variance due to R's processing limitations with our large data set. This is more or less accommodated by the Residual Plots against predicted values and Normal Q-Q Plots.

Additionally, the data provided better lends itself to a multivariate analysis given the large number of variable which is out of the scope of this course.

Next Steps

One might perform a multivariate analysis on the given data while using technology that is able to adequately handle the large amount of data.