Online News Popularity: Regression Analysis

Group 5: Varun Datta (Project Leader), Boxuan Fang (Project Analyst), Yifan Xia (Report Analyst) & Emin Rza (Project Associate)

Supervisor: Dr. So-hee Kang

Introduction

In the ever-evolving digital media landscape, Mashable serves as our case study to delve into the complex dynamics of article virality on social media, primarily measured by share counts. This study aims to unravel the multifaceted factors influencing Mashable articles' success, encompassing content features, multimedia elements, publication timing, and audience engagement. Beyond benefiting content creators and digital marketers with strategic insights, our findings contribute to a broader understanding of content virality in the social media era.

Data Set Introduction :

Our dataset comprises an extensive array of metrics associated with Mashable articles, encompassing 61 variables that include content-specific features (word counts in titles and content, content uniqueness), multimedia elements (image and video counts), metadata characteristics (keyword counts, sentiment indices), and publication timing details (day of the week). Although the original article content is excluded due to copyright constraints, the dataset provides rich statistical data for predicting article share counts. This comprehensive dataset forms the basis for analyzing the multifaceted aspects potentially driving the virality of digital news content.

Research Question:

"What key factors contribute to the virality of Mashable articles in social media, as measured by share count and how can we intepret this relationship in a quantifiable manner?

Hypothesis :

We posit that Mashable article share counts are significantly influenced by a combination of content-related factors, including content length, uniqueness, multimedia elements, publication day, and the sentiment and subjectivity expressed in the article. Our study seeks to validate and quantify these influences, providing actionable insights for content optimization strategies.(Smith & Doe, 2020; Johnson et al., 2021).

Procedure

Step 1: Exploratory Data Analysis (EDA), Data Cleaning, and Checking for Collinearity

Step 2: Stepwise Regression for Main effect Model

Step 3: Development of an Interaction Model

Step 5: Refinement of Interaction Model

Step 6: Final Model Selection and Initial Model Diagnostics

Step 7: Final Model Development using Remedial Methods(if needed)

Step 8: Final Model Validation and Diagnostics

Step 9: Conclusion and Final Assessment

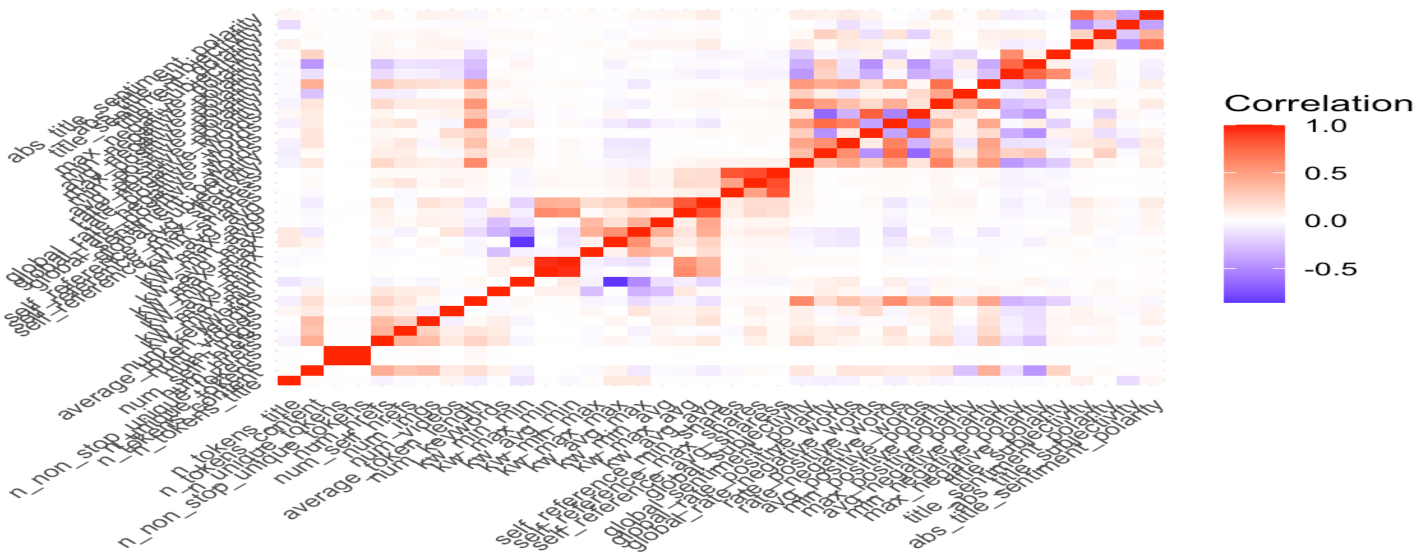
We will be using the following libraries for our analysis
(tidyverse),(reshape2),(plotly),(gridExtra),(MPV),(ggpubr),(olsrr),(lmtest),(webshot2),
(knitr),(MASS),(broom),(caret),(olsrr)

```
## No missing values, no duplicate rows,
##  Single Unique Value Columns: 0 | DataTypes: character, integer, numeric
```

Selecting Predictors from the data set and dataset checks

We will be dropping the weekday specifier columns and the LDA values from the data set due to the nature of these variables and the advice from the prof.We will also be combining the channel type indicator variables into one categorical variable for the purpose of our analysis.

Correlation Plot-HeatMap

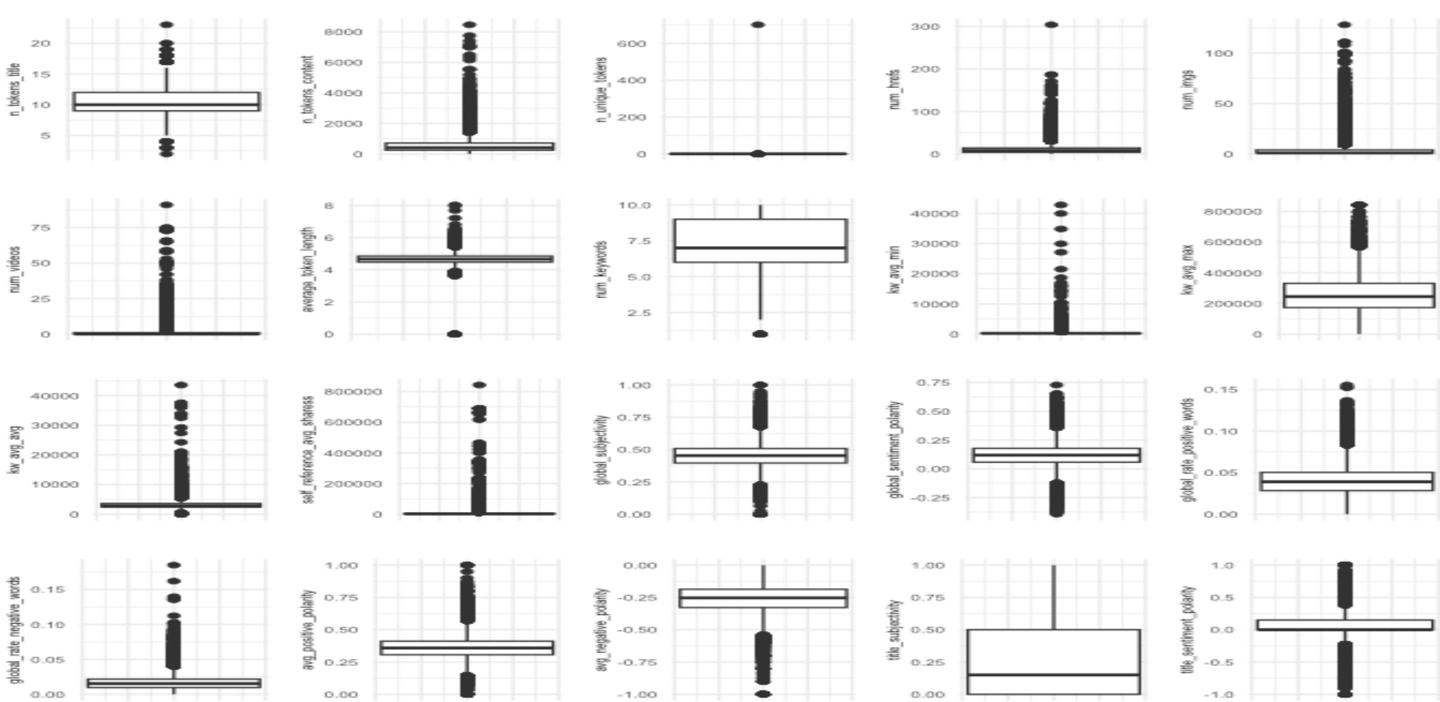


Let's remove highly correlated variables using the plot, most of these variables are max,min values of characteristics which also have an average value or similar metric

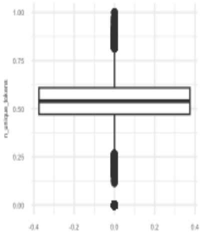
By strategically omitting certain variables, we can ensure a robust model that will be both practical and relevant to content creators and marketers seeking to maximize online engagement.

Description of Dataset

Distributions for each each of the variables we will be analysisng

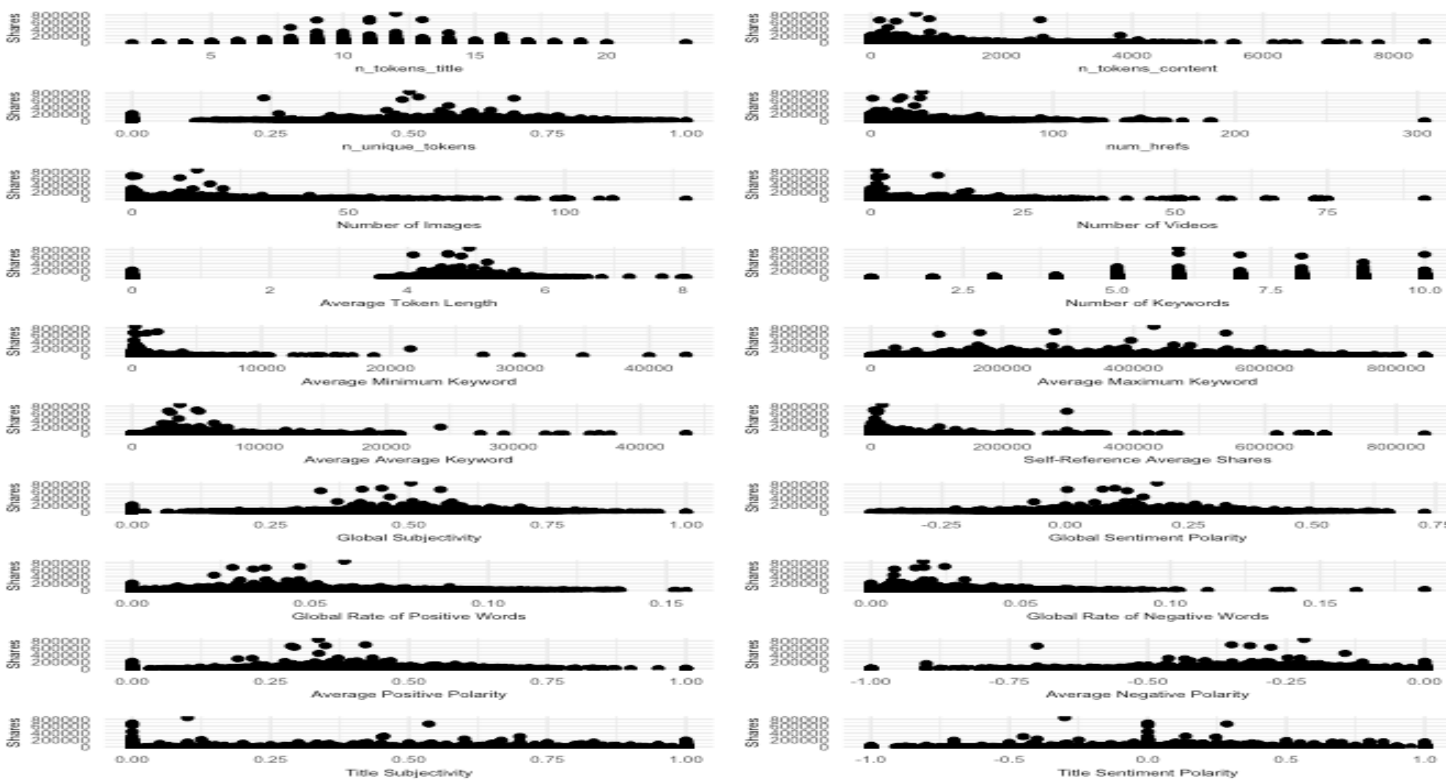


From the boxplot for n_unique_tokens we can see there is a clear outlier, let's rectify that by removing that point



Now we can clearly see the quartiles.

Plots of Shares vs Predictors from our refined dataset



From the scatter plots we can see non of our predictors have a linear relationship with our response indicating that the relationship between these predictors and content shareability is likely more complex than a simple linear model can capture.

Building the Model

We will be using 60% of our data set as our training set and the other 40% as our training set,(view source code)

STEPWISE REGRESSION MODEL SELECTION FOR MAIN EFFECT MODEL

We will use a null model as our lower bound and a model with every single predictor as our upper bound and a null model as our lower bound for a step wise AIC SELECTION with direction = both(view source code).

```
## R^2_ADJ for all predictors in Linear Model:  0.01832238
```

We can definitely increase our R^2 adjusted value and other metrics for this model by simply finding the optimal combination of predictors to use for a good main effect model.

STEP AIC FOR MODEL SELECTION

The Step AIC selection resulted in the following model

Estimates and P-values from Linear Model on the(right)

```
## Adjusted R^2 of this model:  0.01829708
```

From the output we can see the $R^2_{Adjusted}$ value is similar to the full model, so by the principle of parsimony we will proceed with the STEP AIC fit as our model for now.

If we examine the p-values for the t-tests for the significance of the coefficients and their impact on our response variable. We can clearly see from the t-tests that the coefficients of predictors num_keywords and global_rate_positive_words have a p-value 0.05 Let's do a full model reduced Model F-Test to see if we can drop them.

	Estimate	P-value
(Intercept)	1064.6245053	0.0935182
data_channeldata_channel_is_entertainment	-509.2224085	0.0722201
data_channeldata_channel_is_lifestyle	-403.3620540	0.3321925
data_channeldata_channel_is_socmed	113.9015841	0.7705813
data_channeldata_channel_is_tech	-271.6849138	0.3350284
data_channeldata_channel_is_world	-880.1182949	0.0015823
data_channelnot specified	1630.3576118	0.0000004
self_reference_avg_sharess	0.0287073	0.0000000
kw_avg_avg	0.4981239	0.0000000
num_hrefs	24.8756245	0.0008439
avg_negative_polarity	-1457.0251318	0.0441262
is_weekend1	559.5830147	0.0199577
num_keywords	76.0889213	0.0832446
average_token_length	-348.4279713	0.0077444
global_subjectivity	3028.2654808	0.0030746
global_rate_positive_words	-8337.2342688	0.1242542

Step AIC model parameters coefficents with p-value >0.05 for t-test(below)

	Estimate	Pr(>|t|)
(Intercept)	1064.62451	0.0935182
data_channeldata_channel_is_entertainment	-509.22241	0.0722201
data_channeldata_channel_is_lifestyle	-403.36205	0.3321925
data_channeldata_channel_is_socmed	113.90158	0.7705813
data_channeldata_channel_is_tech	-271.68491	0.3350284
num_keywords	76.08892	0.0832446
global_rate_positive_words	-8337.23427	0.1242542

Anova Results for F Test for dropping num_keywords and global_rate_positive_words

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
23772	3647776213724	NA	NA	NA	NA
23770	3646995868203	2	780345521	2.543026	0.0786495

```
##
##  We can drop these two predictors at 0.05 significance level
```

Estimates and P-value for resulting linear model Linear Model

	Estimate	P-value
(Intercept)	1494.8108995	0.0077049
data_channeldata_channel_is_entertainment	-461.2275771	0.1025026
data_channeldata_channel_is_lifestyle	-279.3264814	0.4957275
data_channeldata_channel_is_socmed	89.0030704	0.8196519
data_channeldata_channel_is_tech	-167.5270011	0.5449449
data_channeldata_channel_is_world	-735.0897675	0.0065866
data_channelnot specified	1729.0100479	0.0000001
self_reference_avg_sharess	0.0288176	0.0000000
kw_avg_avg	0.5004625	0.0000000
num_hrefs	27.0494505	0.0002484
avg_negative_polarity	-1502.4662925	0.0371833
is_weekend1	569.2673526	0.0176752
average_token_length	-378.3155329	0.0036331
global_subjectivity	2579.6794435	0.0075880

```
## Adjusted R^2 of this model:  0.01816964
```

This current model is our main effect model

Interaction Model Selection

Let's create a model using our final model and add all possible up to 3 way interactions as our upper bound for a STEP AIC to select the interaction model

Coefficients and P-values for Interaction Model from Step AIC (Rounded to 3 d.p.)

Coefficient	Estimate	P-value
(Intercept)	-489.682	0.505
data_channeldata_channel_is_entertainment	1169.070	0.146
data_channeldata_channel_is_lifestyle	2172.123	0.041
data_channeldata_channel_is_socmed	2740.360	0.003
data_channeldata_channel_is_tech	2068.011	0.017
data_channeldata_channel_is_world	1330.445	0.077
data_channelnot specified	4914.005	0.000
self_reference_avg_sharess	-1.489	0.000
kw_avg_avg	1.157	0.000
num_hrefs	29.533	0.000
avg_negative_polarity	39764.368	0.000
is_weekend1	2579.625	0.006
average_token_length	-432.973	0.003
global_subjectivity	14891.597	0.037
self_reference_avg_sharess:average_token_length	0.325	0.000
self_reference_avg_sharess:avg_negative_polarity	-7.383	0.000
data_channeldata_channel_is_entertainment:self_reference_avg_sharess	-0.004	0.836
data_channeldata_channel_is_lifestyle:self_reference_avg_sharess	-0.044	0.058
data_channeldata_channel_is_socmed:self_reference_avg_sharess	0.003	0.854
data_channeldata_channel_is_tech:self_reference_avg_sharess	-0.004	0.676
data_channeldata_channel_is_world:self_reference_avg_sharess	-0.024	0.051
data_channelnot specified:self_reference_avg_sharess	0.046	0.000
average_token_length:global_subjectivity	-2469.945	0.105
data_channeldata_channel_is_entertainment:kw_avg_avg	-0.507	0.043
data_channeldata_channel_is_lifestyle:kw_avg_avg	-0.697	0.022
data_channeldata_channel_is_socmed:kw_avg_avg	-0.850	0.002
data_channeldata_channel_is_tech:kw_avg_avg	-0.690	0.018
data_channeldata_channel_is_world:kw_avg_avg	-0.687	0.009
data_channelnot specified:kw_avg_avg	-0.986	0.000
is_weekend1:global_subjectivity	-4093.791	0.042
self_reference_avg_sharess:kw_avg_avg	0.000	0.005
avg_negative_polarity:average_token_length	-8670.672	0.000
self_reference_avg_sharess:is_weekend1	-0.037	0.000
self_reference_avg_sharess:avg_negative_polarity:average_token_length	1.573	0.000

Metric	Value
R_Squared	0.046
Adj_R_Squared	0.044
AIC	515235.9
BIC	515518.6
F_Statistic	34.39(33, 23752)
P_Value	0
Sigma Hat	12221.605

Model Summary Statistics (Rounded to 3 d.p.)

Final Interaction Model Our $R^2_{adjusted}$ value has gone up from 0.018 to 0.044(rounded to 3 d.p.) which is a significant improvement. The result of the Global F-Test suggests that our model is significant with a p-value of 0 (rounded to 3 d.p.)

Looking at the p-values here for the t-tests for significance of the coefficients we can clearly drop the following interaction term for average_token_length:global_subjectivity. **We will proceed to drop this term and use the resulting model as our final interaction model**

Note: We are not dropping any other terms as the t tests for at least one of the coefficients related to those categorical variables or their interaction terms is significant.

Summary of Final interaction Model

IM_.sigma	IM_.r.squared	IM_.adj.r.squared
12222.02	0.0455012	0.0442153

Before we proceed forward we are writing a function to compare 2 models using Mallow's CP,PRESS,AIC,BIC and the two coefficient of determination values.(Function Hidden, view source code)

```
##           DELTA_AIC           DELTA_BIC          DELTA_PRESS
##        -620.52532423       -467.06512765 407952427597.73828125
##      DELTA_R_squared   DELTA_Adj_R_squared          Mallows_cp
##           0.02679490          0.02604563         661.79942520
```

From these results we can see that our AIC,BIC went down when comparing Final Interaction Model metrics - Final Main Effect Model metrics, R^2, $R^2_{adjusted}$ went up, which means our interaction model is a better model out of the two and is our choice for the final model.

Interaction Model is our Final Model

Before doing a full diagnostics on the model, let's only check the Homoscedasticity of observed errors assumption from the GAUSS-MARKOV Theorem using a Breusch-Pagan test for the same

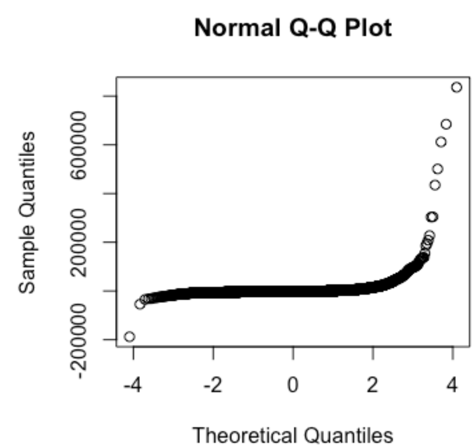
Breusch-Pagan Test Results(rounded to 3 d.p.)

Attribute	Value
Test statistic	465821.6
p-value	0.0
Degrees of freedom	32.0

Prelim Diagnostics

The p-value is close to zero and we can clearly see that the Homoscedasticity of error variance condition is violated.

Let's check the QQ Plot of our residual quantiles with the quantiles of the normal distribution for checking the normality of the residuals.



Our quantiles for the residuals are not the same as the quantiles of the normal distribution, hence normality of residuals/observed errors is violated.

Remidial Transformations

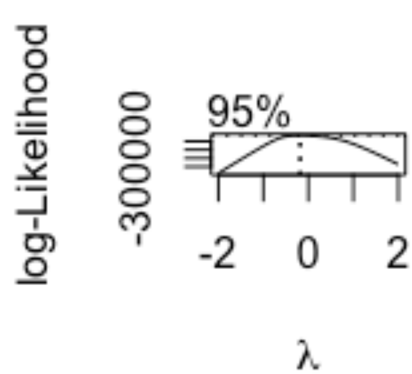
Let's do a box-cox transformation and also apply weighted least squares,subsequently to see if we can rectify the violation of the normality of residuals and the presence of the heteroscedasticity for the residuals.

Box-Cox Transformation

```
## Optimal value of Lambda:  -0.1818182
```

From here the optimal λ is -0.1818182, let's apply the box cox transformation

$Y^* = \frac{Y^{\lambda}-1}{\lambda}$,; if Y is not 0 according to the textbook



Model Summary for Final Model with Box-Cox Transformation(rounded to 3 d.p.)

Metrics	**Values**
R_Squared	**0.112**
Adj_R_Squared	**0.111**
AIC	**-5273.4**
BIC	**-5548**
F_Statistic	**93.883**
P_Value	**0**
Sigma hat	**0.215**

Our Residual Standard Error and MSE have gone down by a lot and we have significantly increase our $R^2_{adjusted}$

Weighted Least Squares Transformation

Using the method from Lecture 22 for case 3 where we use 1/var.s as the weights.(View RMD file for the source code in this section)

Final Model:

Coefficients with p-value > 0.05 :Final Model(below on the right)

	Estimate	PValue
data_channeldata_channel_is_entertainment	-0.0224039	0.1363845
data_channeldata_channel_is_world	-0.0114526	0.4199960
self_reference_avg_sharess	-0.0000027	0.3377822
avg_negative_polarity	-0.0164790	0.8670564
self_reference_avg_sharess:average_token_length	0.0000009	0.1330638
self_reference_avg_sharess:avg_negative_polarity	-0.0000098	0.2183612
data_channeldata_channel_is_lifestyle:self_reference_avg_sharess	0.0000009	0.1092403
data_channeldata_channel_is_tech:self_reference_avg_sharess	-0.0000003	0.1603435
data_channeldata_channel_is_world:self_reference_avg_sharess	0.0000003	0.3586415
data_channelnot specified:self_reference_avg_sharess	0.0000004	0.0695236
data_channeldata_channel_is_tech:kw_avg_avg	-0.0000023	0.7030874
data_channeldata_channel_is_world:kw_avg_avg	-0.0000095	0.0668030
is_weekend1:global_subjectivity	-0.0382945	0.3278283
avg_negative_polarity:average_token_length	-0.0006744	0.9743709
self_reference_avg_sharess:avg_negative_polarity:average_token_length	0.0000023	0.1811259

From the table above we can drop **self_reference_avg_sharess:avg_negative_polarity:average_token_length** from our model using the results of the t-tests.

After dropping the 3-way interaction term, let's use a full model reduced model F Test to see which parameters can we drop with the table above us as our guide.

Anova Results from Full Model Reduced Model F Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
23759	1969.386	NA	NA	NA	NA
23754	1968.894	5	0.4922517	1.187768	0.3122519

With a p-value of 0.3122519 from the Full Model-Reduced Model F-Test , we can drop the following terms self_reference_avg_sharess:average_token_length , self_reference_avg_sharess:avg_negative_polarity ,avg_negative_polarity:average_token_length ,is_weekend:global_subjectivity

$$\begin{aligned}
\frac{\widehat{shares^{\lambda}-1}}{\lambda} =& \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathrm{d.c._ent} + \hat{\beta}_2 \cdot \mathrm{d.c._life} + \hat{\beta}_3 \cdot \mathrm{d.c._socmed} + \hat{\beta}_4 \cdot \mathrm{d.c._tech} \\
&+ \hat{\beta}_5 \cdot \mathrm{d.c._world} + \hat{\beta}_6 \cdot \mathrm{d.c._not_spec} + \hat{\beta}_7 \cdot \mathrm{sr_as} + \hat{\beta}_8 \cdot \mathrm{kw_aa} + \hat{\beta}_9 \cdot \mathrm{num_hrefs} \\
&+ \hat{\beta}_{10} \cdot \mathrm{is_weekend} + \hat{\beta}_{11} \cdot \mathrm{avg_tok_len} + \hat{\beta}_{12} \cdot \mathrm{glob_subj} \\
&+ \hat{\beta}_{13} \cdot \mathrm{d.c._ent} : \mathrm{sr_as} + \hat{\beta}_{14} \cdot \mathrm{d.c._life} : \mathrm{sr_as} + \hat{\beta}_{15} \cdot \mathrm{d.c._socmed} : \mathrm{sr_as} \\
&+ \hat{\beta}_{16} \cdot \mathrm{d.c._tech} : \mathrm{sr_as} + \hat{\beta}_{17} \cdot \mathrm{d.c._world} : \mathrm{sr_as} + \hat{\beta}_{18} \cdot \mathrm{d.c._not_spec} : \mathrm{sr_as} \\
&+ \hat{\beta}_{19} \cdot \mathrm{d.c._ent} : \mathrm{kw_aa} + \hat{\beta}_{20} \cdot \mathrm{d.c._life} : \mathrm{kw_aa} + \hat{\beta}_{21} \cdot \mathrm{d.c._socmed} : \mathrm{kw_aa} \\
&+ \hat{\beta}_{22} \cdot \mathrm{d.c._tech} : \mathrm{kw_aa} + \hat{\beta}_{23} \cdot \mathrm{d.c._world} : \mathrm{kw_aa} + \hat{\beta}_{24} \cdot \mathrm{d.c._not_spec} : \mathrm{kw_aa} \\
&+ \hat{\beta}_{25} \cdot \mathrm{sr_as} : \mathrm{kw_aa} + \hat{\beta}_{26} \cdot \mathrm{sr_as} : \mathrm{is_weekend}
\end{aligned}$$

FINAL MODEL: EQUATION AND INTERPRETATION

Abbreviation Dictionary: D.C.: Data Channel,S.R.A.S.: Self Reference Avg Shares , K.A.A.: Keyword Avg Avg, N.H.: Num Hrefs ,I.W.: Is Weekend, A.T.L.: Average Token Length, G.S.: Global Subjectivity

Interpretation of our model's coefficients and their significance

Coefficient	Variable	Value	Coefficient	Variable	Value
$\hat{\beta}_0$	Intercept	3.9247237704505	$\hat{\beta}_1$	d.c._ent	-0.0215401553683
$\hat{\beta}_2$	d.c._life	0.1181528563169	$\hat{\beta}_3$	d.c._socmed	0.1678115350934
$\hat{\beta}_4$	d.c._tech	0.0599418855394	$\hat{\beta}_5$	d.c._world	-0.0108433698279
$\hat{\beta}_6$	d.c._not_spec	0.0925455391235	$\hat{\beta}_7$	sr_as	0.0000011416165
$\hat{\beta}_8$	kw_aa	0.0000493694856	$\hat{\beta}_9$	num_hrefs	0.0014029322548
$\hat{\beta}_{10}$	is_weekend	0.0817229941257	$\hat{\beta}_{11}$	avg_tok_len	-0.0213240880211
$\hat{\beta}_{12}$	glob_subj	0.1315670052225	$\hat{\beta}_{13}$	d.c._ent : sr_as	0.0000018100158
$\hat{\beta}_{14}$	d.c._life : sr_as	0.0000008879054	$\hat{\beta}_{15}$	d.c._socmed : sr_as	0.0000012351376
$\hat{\beta}_{16}$	d.c._tech : sr_as	-0.0000002254488	$\hat{\beta}_{17}$	d.c._world : sr_as	0.0000004401177
$\hat{\beta}_{18}$	d.c._not_spec : sr_as	0.0000004303885	$\hat{\beta}_{19}$	d.c._ent : kw_aa	-0.0000111991485
$\hat{\beta}_{20}$	d.c._life : kw_aa	-0.0000340864494	$\hat{\beta}_{21}$	d.c._socmed : kw_aa	-0.0000333004224
$\hat{\beta}_{22}$	d.c._tech : kw_aa	-0.0000022513886	$\hat{\beta}_{23}$	d.c._world : kw_aa	-0.0000097514656
$\hat{\beta}_{24}$	d.c._not_spec : kw_aa	-0.0000225867695	$\hat{\beta}_{25}$	sr_as : kw_aa	-0.0000000001121
$\hat{\beta}_{26}$	sr_as : is_weekend	-0.0000009101439			

Coefficients with p-value > 0.05 for Overall Final Model

	Estimate	PValue
data_channeldata_channel_is_entertainment	-0.0215402	0.1513230
data_channeldata_channel_is_world	-0.0108434	0.4440445
data_channeldata_channel_is_lifestyle:self_reference_avg_sharess	0.0000009	0.1005831
data_channeldata_channel_is_tech:self_reference_avg_sharess	-0.0000002	0.2856426
data_channeldata_channel_is_world:self_reference_avg_sharess	0.0000004	0.1005927
data_channelnot specified:self_reference_avg_sharess	0.0000004	0.0742005
data_channeldata_channel_is_tech:kw_avg_avg	-0.0000023	0.7090329
data_channeldata_channel_is_world:kw_avg_avg	-0.0000098	0.0590008

$\hat{\beta}_0$ The transformed value of shares, when it is a weekend day(reference level or zero value for is_weekend),the channel is business(reference category for data_channel), while all the other predictors are 0. $\hat{\beta}_1$ *to* $\hat{\beta}_6$: The difference in transformed value of shares, when comparing data channel for the respective coefficients channel name with the business data channel, holding all the other predictors constant. This value has now real world meaning as there will be no scenario where some of our predictors could actually be zero in an article.

$\hat{\beta}_{10}$: The difference in transformed value of shares for when a weekday is compared with a weekend day holding all other predictors constant. From the table above us we can see there is no statistically significant pairwise difference between business and entertainment & world and business data channels.

$\boldsymbol{\hat{\beta}_7}$ ***to*** $\boldsymbol{\hat{\beta}_9}$ ***and*** $\boldsymbol{\hat{\beta}_9}$ ***to***$\boldsymbol{\hat{\beta}_{12}}$:The change in transformed value of shares for one unit change in the respective variable of the coefficient, holding all other predictors constant.

$\hat{\beta}_{13}$ *to* $\hat{\beta}_{18}$:These terms represent the interaction between different data channels (like entertainment, life, etc.) and the self-reference average shares (S.R.A.S.). Each coefficient reflects the difference in the change of transformed shares per one unit change in S.R.A.S., compared to the business channel which serves as the reference level holding other predictors constant. A negative values means the response goes down and vice-versa. From the table above, the interaction effect due to channels lifestyle,tech, not specified and world are not statistically significant.

$\hat{\beta}_{19}$ *to* $\hat{\beta}_{24}$: These terms represent the interaction between different data channels (like entertainment, life, etc.) and Key-Words-Average -kw_avg_avg- . Each coefficient reflects the difference in the change of transformed shares per one unit change in K.A.A (while holding other predictors constant), compared to the business channel which serves as the reference level . A negative values means the response goes down and vice-versa. The interaction effect due to the not specified and tech channels are not statistically significant according to the table above.

$\hat{\beta}_{25}$: The combined effect of Keyword Avg Avg and Self Reference Avg Shares on the transformed shares. It represents how the effect of one unit increase in 'kw_avg_avg' on the dependent variable 'transformed_shares' changes for each unit increase in 'self_reference_avg_sharess' while holding all else constant. A negative Beta26 indicates that as 'self_reference_avg_sharess' increases, the positive effect of 'kw_avg_avg' on 'transformed_shares' decreases."

$\hat{\beta}_{26}$:When comparing an article published on a weekend vs weekday, we have a -0.0000009101439 difference in the transformed shares for each one unit change in 'self_reference_avg_sharess' while holding all the other predictors constant.

Model Metrics and interpretation

Final Model Metrics (rounded to 3 d.p.)

	MSE	R_Squared	Adj_R_Squared	AIC	BIC	PRESS	F_Statistic	F_pvalue
value	0.083	0.116	0.115	-5747.97	-5521.818	1111.156	119.5656	0

Adjusted R^2: From this we can see that approximately 11.5% of the variation in our response variable is explained by the regression model after adjusting for our parameters.Our model has a low explanatory power. **F-Test:** The extremely low p-value of the global F-test for our linear model suggests a statistically significant association between the predictors and the response variable, indicating that the model as a whole is likely to be meaningful. **MSE:** With a value of 0.083, it indicates the average squared difference between the observed actual outcomes and the outcomes predicted by the model is very low.

Model Validation

Using the 40% of our intitial data set for testing the preidciton capabilities of the model.

```
## MSPE:  0.03731802   & MSE:  0.08289012
```

Our model's lower Mean Squared Prediction Error (MSPE) compared to its Mean Squared Error (MSE) suggests better performance on unseen data, indicating potential for good generalizability and lack of overfitting. To confirm this, we should employ k-fold cross-validation, a robust validation technique. In this process, the data is divided into 'k' parts; the model is trained on 'k-1' folds and tested on the remaining fold, iteratively. This will provide a more comprehensive assessment of the model's generalization ability across various subsets of the data.

K-Folds Cross Validation

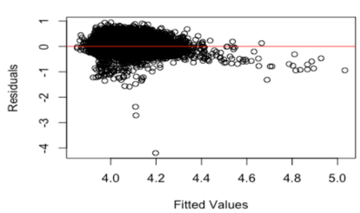
We will do a K-Folds Cross Validation Technique with 100 folds, typically used in machine learning. You take a model, split the dataset into K sections, train it on 1 section and test on K-1 sections and repeat it for all the folds to get an average MSE/AIC etc.

```
## AVG K FOLDS MSPE:  0.04636506
```

For a 100 fold CV, we can see our MSPE Average is much lower than our MSE, which confirms what we had mentioned about generalizability of our model.

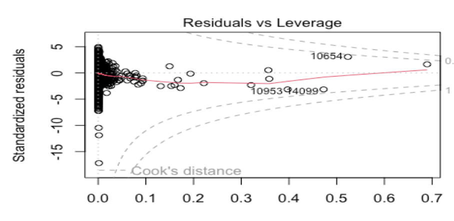
Diagnostics

Diagnostic Plots and Outliers



Residual vs Fitted Plot :

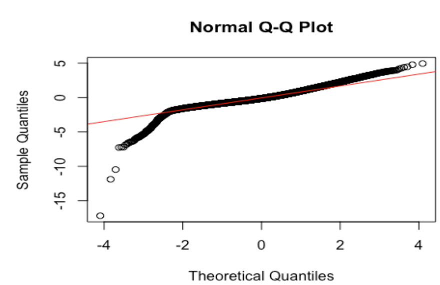
In this plot, the residuals do not appear to fan out or form a pattern, which is good for homoscedasticity. However, there's a slight curve to the residual points, which may suggest a non-linear relationship between the predictors and the response variable.



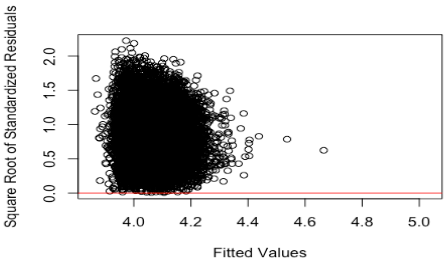
Residuals vs. Leverage Plot:

The plot shows Cook's distance as dashed curves, which measures the influence of each observation. In this plot, there are a few points labeled that are well outside the Cook's distance curves, especially the one labeled 106540, indicating they are potentially influential.

QQ Normal Plot



The curvature in this Q-Q plot suggests that the residuals have heavier tails than expected under normality. This means that there are more extreme values (both low and high) than what would be expected if the residuals were perfectly normally distributed. This could affect confidence intervals and hypothesis tests, as these are typically based on the assumption of normally distributed errors.



Scale-Location Plot

The red line should that should ideally be horizontal and flat across the range of fitted values if homoscedasticity holds. In our plot, the loess fit line shows a slight upward trend, suggesting that the variance of the residuals may increase as the fitted values increase, which indicates potential heteroscedasticity. The points also seem to fan out slightly for higher fitted values.

OUTLIERS AND INFLUENTIAL POINTS FINDING THEM USING R FUNCTIONS

We are using the R functions to retrieve a list of points which fail the threshold of our measures like DFFBETAS, COOK's Distance,DFFITS and Leverage.

```
## Total number of inflential points using thresholds for leverage, cooks distance and dffits and hat 1557
```

This is a huge chunk of our data set deleting these points might improve our metrics but we cannot be sure about the reliability.

Given the lack of ownership and detailed background knowledge of the dataset, coupled with the limited scope of this study, we will refrain from removing influential points or outliers. Such a procedure, without a thorough understanding of the underlying data-generating process, risks the exclusion of a significant portion of the dataset. This could potentially lead to overfitting and adversely affect the model's generalizability to the broader population. Therefore, to maintain the integrity and applicability of our findings, all data points will be retained for analysis.

CONCLUSION

Summary of Findings:

Our comprehensive statistical analysis has identified critical factors influencing the shareability of Mashable articles. The model highlights the intricate relationships between variables such as data channel types and their interactions with Self-Reference Average Shares (S.R.A.S.) and Keyword Average (K.A.A.), which differently affect share counts. It underscores the significance of the data channel type and its interplay with S.R.A.S., while noting that other interactions, like those with keywords, are less influential. The analysis also shows variations in share counts based on publication day and data channel, providing valuable insights for content optimization strategies.

Limitations of Our Study:

Variable Selection: Important factors like social media algorithms and real-time events were not included, possibly leading to unaccounted influences on shareability. -Model Complexity: While our model is comprehensive, its complexity might mask simpler relationships. Data Quality and Accuracy: Assumptions about data accuracy and completeness might have impacted our findings. -Need for Alternative Modeling Approaches: The low R^2 value (0.111), minimal mean squared error, unequal variances in observed errors, violation of normality, and presence of influential points and outliers in our dataset suggest limitations in our linear regression approach. This indicates the necessity for alternative, more robust modeling techniques.

Future Extensions:

To address these challenges, future studies could leverage advanced machine learning algorithms like Support Vector Machines (SVMs) and Random Forests. These methods, renowned for their ability to handle complex, high-dimensional data, and large datasets with numerous input variables, respectively, could offer more sophisticated insights into the multifaceted predictors of article shareability. By employing these techniques, we anticipate a more robust and nuanced understanding of the dynamics shaping online content virality.

REFERENCES

Fernandes,Kelwin, Vinagre,Pedro, Cortez,Paulo, and Sernadela,Pedro. (2015). *Online News Popularity*. UCI Machine. Learning Repository. https://doi.org/10.24432/C5NS3V.