Group B12: Sarah Yu, Sophia Zhou, Daniel Zhu, Jennifer Zwiebel
Professor Vishwanath
Math 189
9 June 2024

Math 189 Final Project

Introduction to the project

Our final project explores which attributes of online articles are correlated with a higher number of shares in social networks. This problem is relevant because online news is so prevalent today. The spread of news and public awareness of issues are of great importance in our world. We hope that this project helps us gain insight into the article attributes that are correlated with more article shares. Additionally, the results from this project can provide strategies for online article publishers to create more widely shared articles.

Our data was obtained from the UC Irvine Machine Learning Repository Online News Repository dataset. This data comes from articles published by Mashable, a digital media site. The dataset includes the dependent variable “shares” and 60 other columns with possible covariates that measure a range of metrics such as keywords, number of visual media, day of the week published, and much more. In total, this dataset contains 39,644 observations.

We found two papers that already used this dataset. One article, “A Novel Family of Boosted Online Regression Algorithms with Strong Theoretical Bounds” by Kari, Khan, Ciftci, and Kozat (2016), explored different types of regression algorithms. They investigated broader statistical methods regarding mean squared errors (MSE), which is quite different from our question. In the other article, “Optimal Sub-sampling with Influence Functions” by Ting and Brochu (2017), researchers were interested in general statistical methods and the problem of optimal sub-sampling, rather than specifically analyzing the data we used.

Our first step in our analysis was to clean the data. We removed the spaces before the names of the columns. Additionally, we put the weekdays and data channels into their own respective variables. Then, we fitted the full linear regression model with all the variables. Although the R-squared value is low, our p-value for our F-statistic is miniscule, which indicates that the model is indeed a good fit for the data.

R-squared:	0.022
Adj. R-squared:	0.020
F-statistic:	13.68
Prob (F-statistic):	1.31e-121
Log-Likelihood:	-3.5399e+05
AIC:	7.081e+05
BIC:	7.086e+05

Figure (1) shows part of the output of the summary of the full linear regression model using all of the variables.

We also plotted a histogram of our dependent variable to get a general sense of what the data looks like:

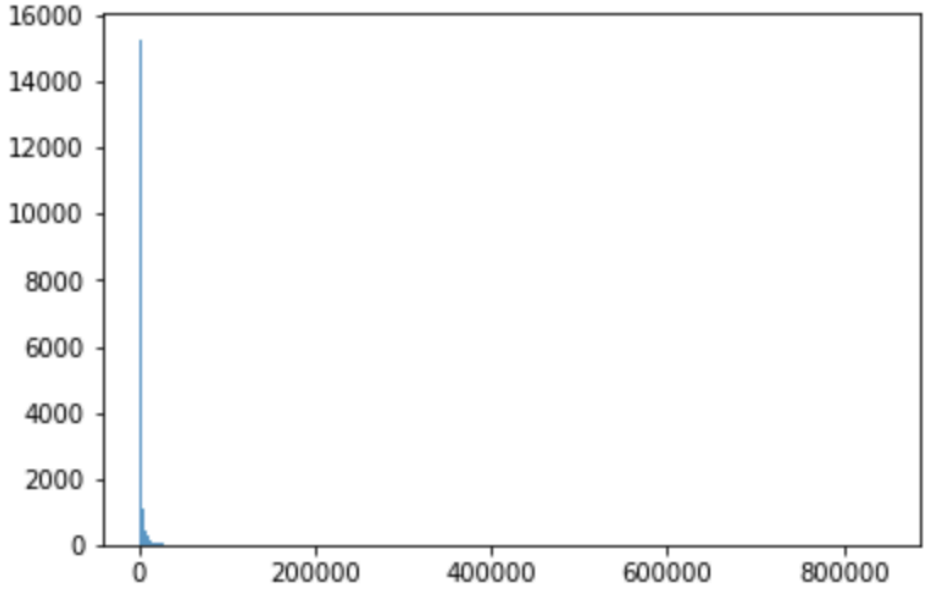


Figure (2). This shows the histogram of the dependent variable 'shares'. As we can see, the data is highly skewed. The bulk of the data lies below 30,000, but there are a few large outliers.

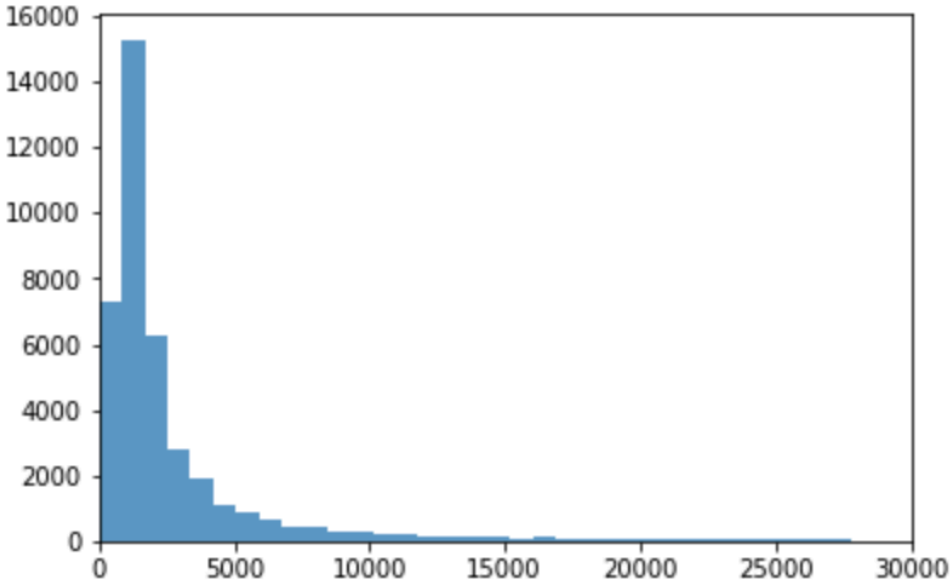


Figure (3) shows a zoomed in version of the histogram above. The data appears to roughly follow an exponential decay.

Variable selection

Next, we used the LASSO variable selection method to select the most relevant variables from this dataset. First, we needed to figure out which penalty parameter to use. To figure this out, we plotted the variables under the LASSO model against different penalty parameters, as seen below:

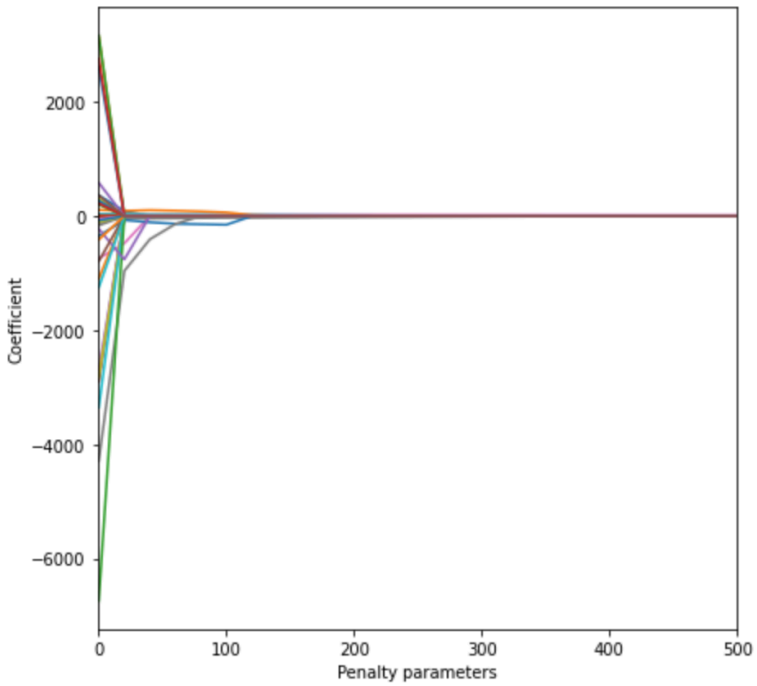


Figure (4) shows the LASSO regularization path of all of the variables.

Based on this graph, we plotted a more zoomed-in version:

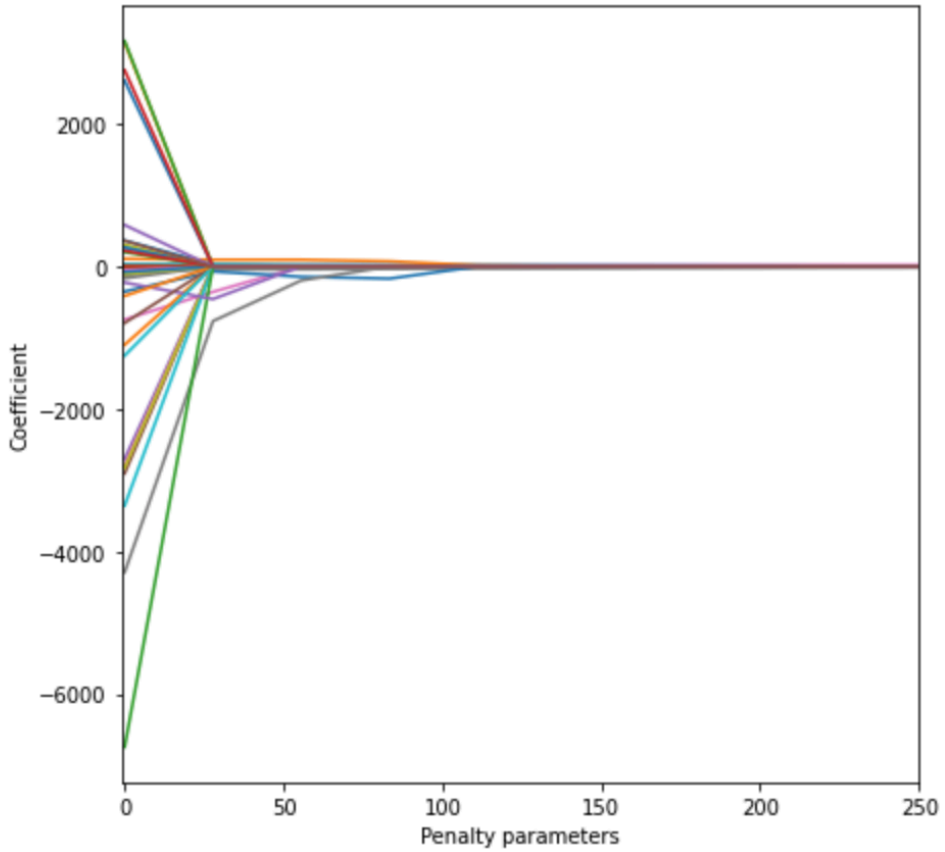


Figure (5) shows a zoomed-in version of the LASSO regularization path. Since the majority of the variables appear to converge to 0 just around 150, we chose this as our penalty parameter.

Based on this, we chose a penalty parameter of 150. After using the LASSO model with this penalty parameter, we were left with 16 variables: 'num_self_hrefs', 'num_imgs', 'num_videos', 'num_keywords', 'kw_min_min', 'kw_max_min', 'kw_avg_min', 'kw_min_max', 'kw_max_max', 'kw_avg_max', 'kw_min_avg', 'kw_max_avg', 'kw_avg_avg', 'self_reference_min_shares', 'self_reference_max_shares', 'self_reference_avg_sharess'.

The next step was to check for multicollinearity between these selected variables. To do this, we looked at the VIFs of all of the selected variables. We can also visualize this with the correlation coefficients.

```
VIF: num_self_hrefs:  1.160
VIF: num_imgs:  1.132
VIF: num_videos:  1.064
VIF: num_keywords:  1.381
VIF: kw_min_min:  3.817
VIF: kw_max_min:  10.973
VIF: kw_avg_min:  10.622
VIF: kw_min_max:  1.347
VIF: kw_max_max:  4.300
VIF: kw_avg_max:  2.962
VIF: kw_min_avg:  2.040
VIF: kw_max_avg:  5.849
VIF: kw_avg_avg:  7.325
VIF: self_reference_min_shares:  6.596
VIF: self_reference_max_shares:  8.370
VIF: self_reference_avg_sharess:  19.048
```

Figure (6) shows the VIFs of the seventeen variables that resulted from the LASSO model. Some of the VIFs are greater than 5, so we removed some highly correlated variables.

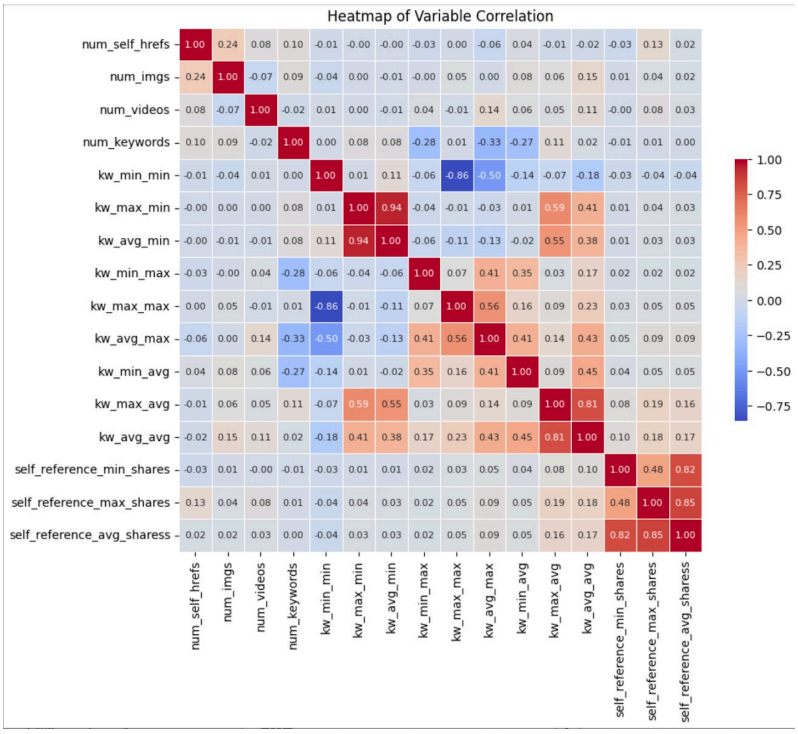


Figure (7) shows a heatmap of the correlation coefficients between the seventeen variables.

Once we removed variables that had high multicollinearity, we were left with 13 variables: 'num_self_hrefs', 'num_imgs', 'num_videos', 'num_keywords', 'kw_min_min', 'kw_avg_min', 'kw_min_max', 'kw_max_max', 'kw_avg_max', 'kw_min_avg', 'kw_max_avg', 'self_reference_min_shares', and 'self_reference_max_shares'. The VIF for all of these remaining variables was below 5, as seen in the screenshot below:

```
VIF: num_self_hrefs:  1.122
VIF: num_imgs:  1.091
VIF: num_videos:  1.056
VIF: num_keywords:  1.297
VIF: kw_min_min:  3.793
VIF: kw_avg_min:  1.543
VIF: kw_min_max:  1.336
VIF: kw_max_max:  4.295
VIF: kw_avg_max:  2.372
VIF: kw_min_avg:  1.309
VIF: kw_max_avg:  1.645
VIF: self_reference_min_shares:  1.319
VIF: self_reference_max_shares:  1.396
```

Figure (8) shows the VIFs for the remaining variables. Now, all of the VIFs are below 5, so there is no significant multicollinearity affecting the model.

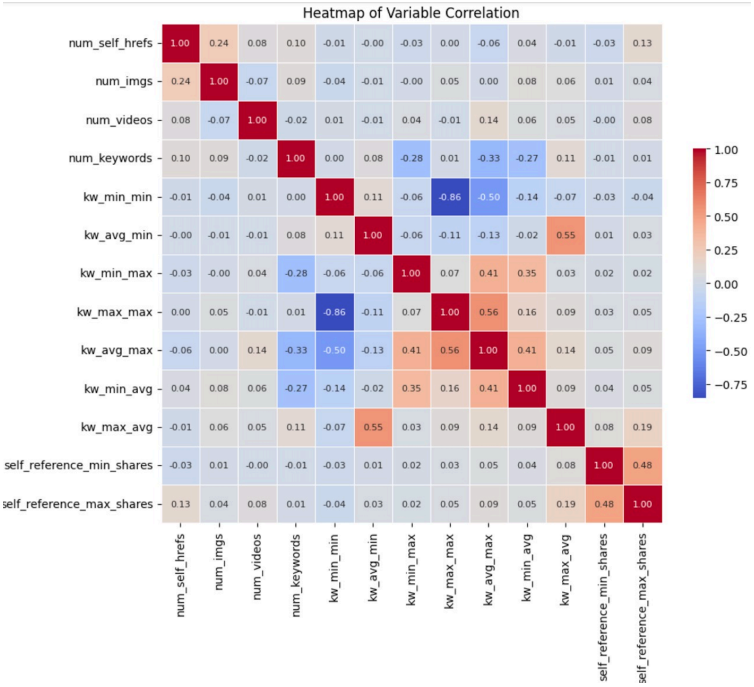


Figure (9) shows the heatmap for the remaining 13 variables. Now, the correlation coefficients between variables are more reasonable.

Regression assumptions

These 13 variables were our final variables that we included in the model. Before we made the linear model with these variables, we needed to check the assumptions for running a linear regression. Our work in checking the assumptions can be seen below:

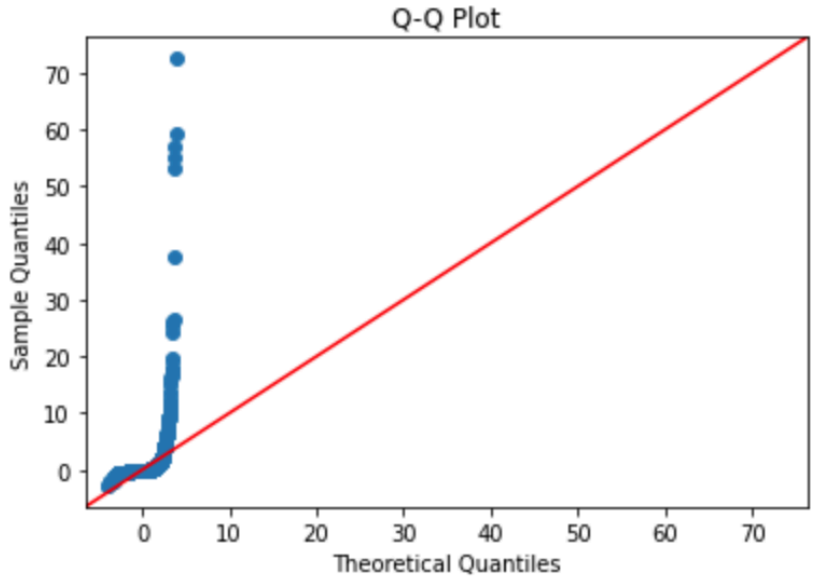


Figure (10) shows the QQ-plot between the data and the normal distributions. Clearly, the assumption of normality is not met.

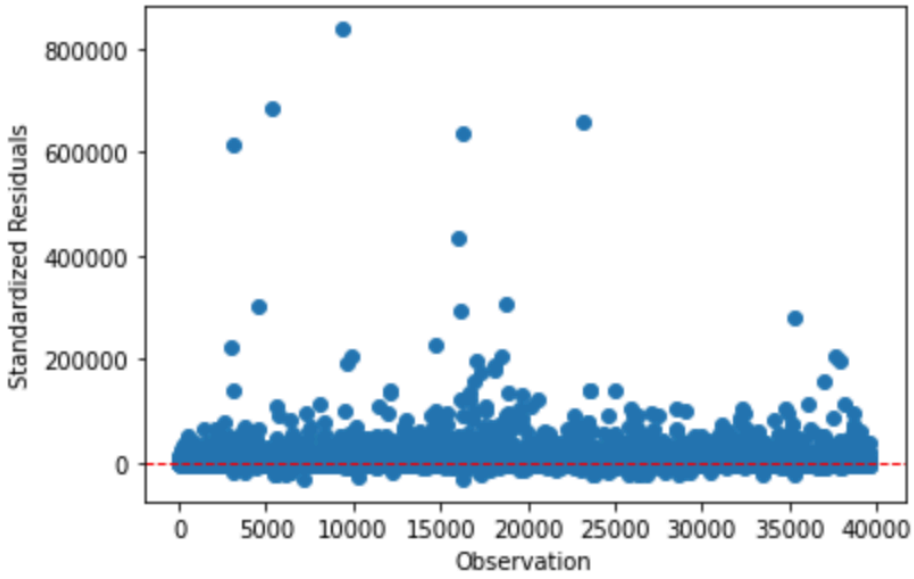


Figure (11) shows the scatterplot of the standardized residuals. This shows that the assumption of independence is not met.

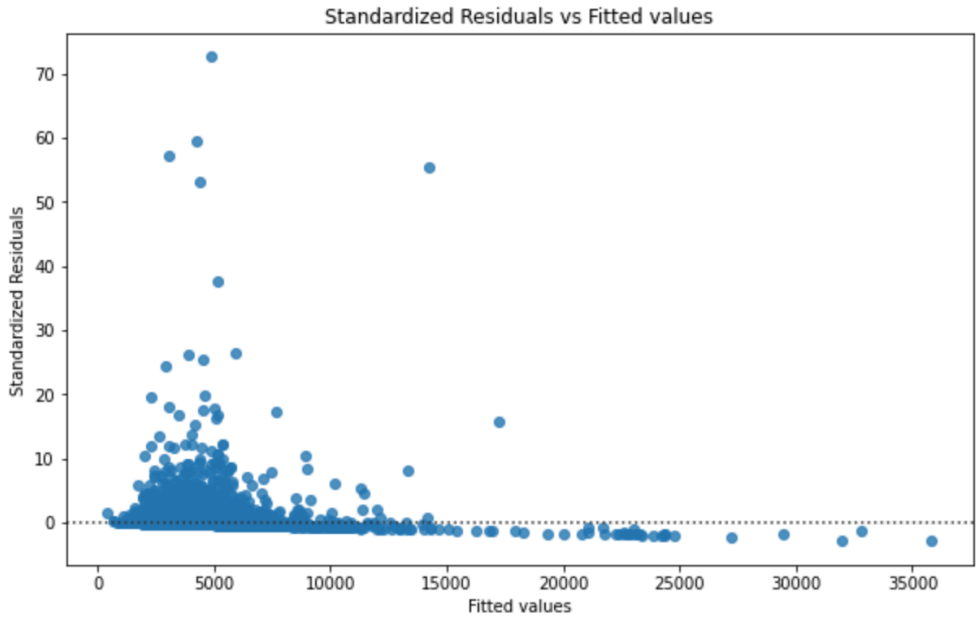


Figure (12) shows the graph of the standardized residuals vs. the dependent variable. This shows that the assumption of linearity is not met.

Finally, we ran the Breusche-Pagan test and got a value of $3.277 * 10^{-8}$. So, we reject the null hypothesis of homoscedasticity.

Assumptions for log-transformed data

As you can see, the assumptions were definitely not met. So, we did a log transform on our dependent variable “shares”. We ran the assumptions again with the log transformed model, and they looked much more reasonable, however they were still not fully met.

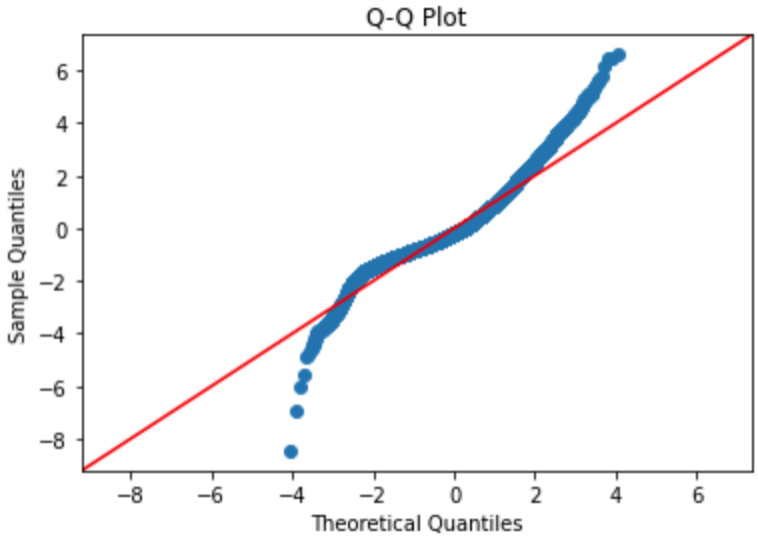


Figure (13) shows the QQ-plot between the data and the normal distributions. Now, the QQ-plot appears to more closely follow the line y=x, and the assumption of normality is more reasonable.

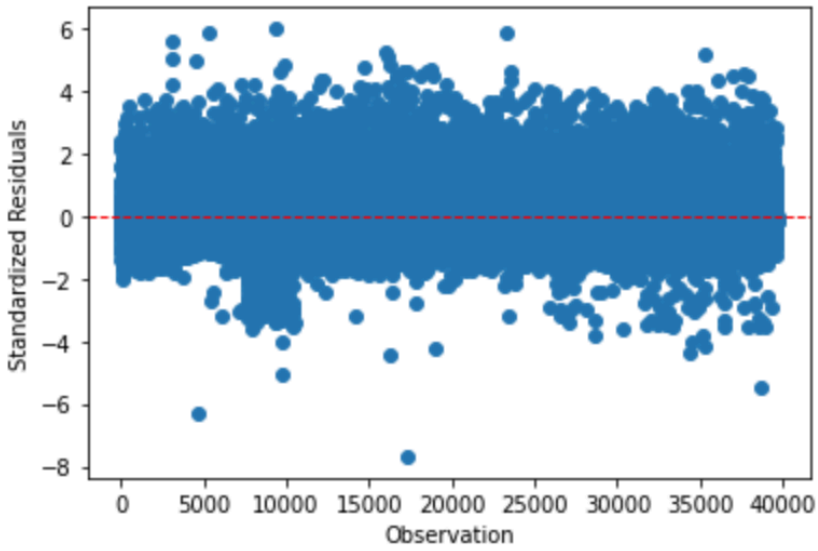


Figure (14) shows the scatterplot of the standardized residuals. Now, since these appear to be randomly scattered around 0, the assumption of independence is clearly met.

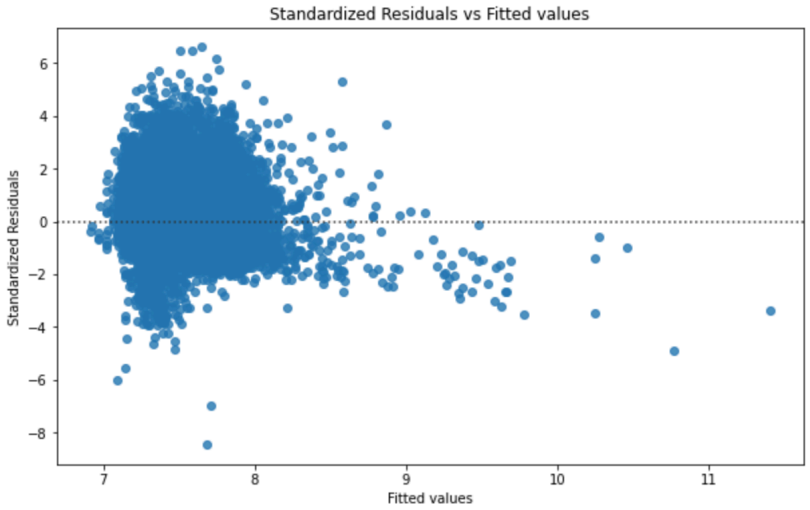


Figure (15) shows the graph of the standardized residuals vs. the dependent variable. This graph has improved, but much of the data is still concentrated to the left side. Thus, we cannot say that the assumption of linearity is met.

Lastly, we ran the Breusche-Pagan test and got a value of $9.796 * 10^{-163}$. So, again, we reject the null hypothesis of homoscedasticity.

Final model

OLS Regression Results

Dep. Variable:	np.log(shares)	R-squared:	0.046
Model:	OLS	Adj. R-squared:	0.046
Method:	Least Squares	F-statistic:	147.4
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	0.00
Time:	15:32:07	Log-Likelihood:	-52460.
No. Observations:	39644	AIC:	1.049e+05
Df Residuals:	39630	BIC:	1.051e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.7081	0.040	169.357	0.000	6.630	6.786
num_self_hrefs	0.0004	0.001	0.348	0.728	-0.002	0.003
num_imgs	0.0081	0.001	14.106	0.000	0.007	0.009
num_videos	0.0034	0.001	2.977	0.003	0.001	0.006
num_keywords	0.0497	0.003	18.264	0.000	0.044	0.055
kw_min_min	0.0012	0.000	9.788	0.000	0.001	0.001
kw_avg_min	-1.039e-05	9.13e-06	-1.138	0.255	-2.83e-05	7.51e-06
kw_min_max	-5.81e-07	9.1e-08	-6.385	0.000	-7.59e-07	-4.03e-07
kw_max_max	1.058e-09	4.41e-08	0.024	0.981	-8.54e-08	8.75e-08
kw_avg_max	6.441e-07	5.2e-08	12.378	0.000	5.42e-07	7.46e-07
kw_min_avg	8.772e-05	4.59e-06	19.100	0.000	7.87e-05	9.67e-05
kw_max_avg	1.101e-05	9.6e-07	11.464	0.000	9.12e-06	1.29e-05
self_reference_min_shares	2.328e-06	2.66e-07	8.767	0.000	1.81e-06	2.85e-06
self_reference_max_shares	5.916e-07	1.31e-07	4.500	0.000	3.34e-07	8.49e-07

Omnibus:	6691.865	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15737.859
Skew:	0.966	Prob(JB):	0.00
Kurtosis:	5.408	Cond. No.	7.19e+06

Figure (16) shows the summary of the regression model having log transformed the dependent variable.

Based on this log-transformed linear regression, the interpretation of the coefficients is:

- Holding all other variables constant, if we increase the variable num_self_hrefs by one unit, the response variable shares will increase, on average, by a factor of $e^{0.0004}$ units. However, it should also be noted that the p-value for this coefficient is 0.728 > 0.05. Therefore, we do not have statistical evidence that the coefficient is different from 0.

- Holding all other variables constant, if we increase the variable num_imgs by one unit, the response variable shares will increase, on average, by a factor of $e^{0.0081}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable num_videos by one unit, the response variable shares will increase, on average, by a factor of $e^{0.0034}$ units. The p-value for this coefficient is 0.03 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable num_keywords by one unit, the response variable shares will increase, on average, by a factor of $e^{0.0497}$ units.

The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable kw_min_min by one unit, the response variable shares will increase, on average, by a factor of $e^{0.0012}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable kw_avg_min by one unit, the response variable shares will increase, on average, by a factor of $e^{-1.039e-05}$ units. However, it should also be noted that the p-value for this coefficient is 0.255 > 0.05. Therefore, we do not have statistical evidence that the coefficient is different from 0.

- Holding all other variables constant, if we increase the variable kw_min_max by one unit, the response variable shares will increase, on average, by a factor of $e^{-5.81e-07}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable kw_max_max by one unit, the response variable shares will increase, on average, by a factor of $e^{1.058e-09}$ units. However, it should also be noted that the p-value for this coefficient is 0.981 > 0.05. Therefore, we do not have statistical evidence that the coefficient is different from 0.

- Holding all other variables constant, if we increase the variable kw_avg_max by one unit, the response variable shares will increase, on average, by a factor of e $e^{6.441e-07}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable kw_min_avg by one unit, the response variable shares will increase, on average, by a factor of $e^{8.772e-05}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable kw_max_avg by one unit, the response variable shares will increase, on average, by a factor of $e^{1.101e-05}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable self_reference_min_shares by one unit, the response variable shares will increase, on average, by a factor of $e^{2.328e-06}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

- Holding all other variables constant, if we increase the variable self_reference_max_shares by one unit, the response variable shares will increase, on average, by a factor of $e^{5.916e-07}$ units. The p-value for this coefficient is 0.000 < 0.05, indicating that this coefficient is indeed different than 0.

Exploratory Analysis – Adding Interaction Terms

However, two of the assumptions – linearity and homoscedasticity – were still not met, so we tried to add all of 78 possible interaction terms to improve these. We used combinations to generate all two-way interaction terms, and included them in the model. We then checked the assumptions of this model.

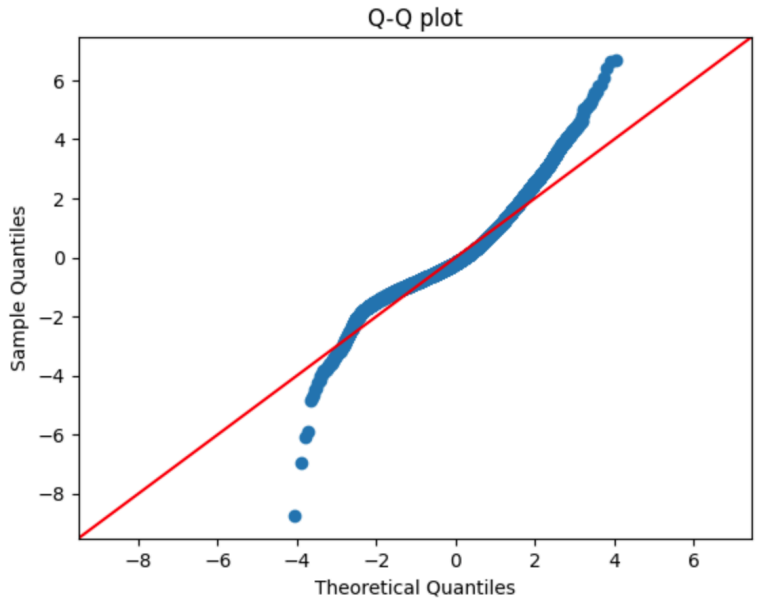


Figure (17) shows the QQ-plot between the data and the normal distributions. The QQ-plot appears to follow the line y=x, and the assumption of normality is reasonable.

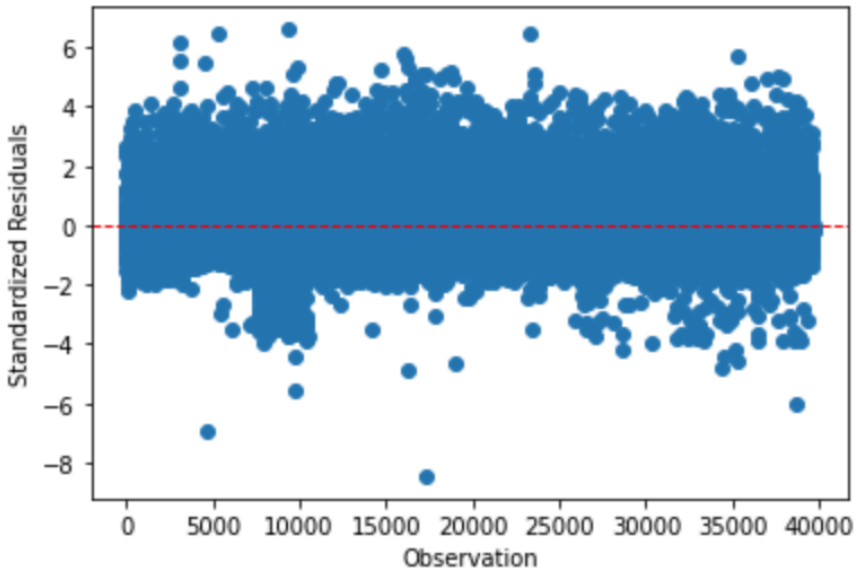


Figure (18) shows the scatterplot of the standardized residuals. Since these appear to be randomly scattered around 0, the assumption of independence is clearly met.

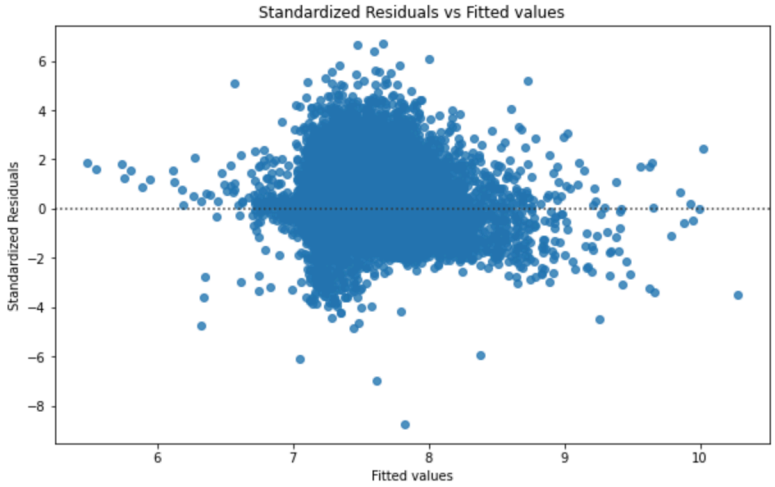


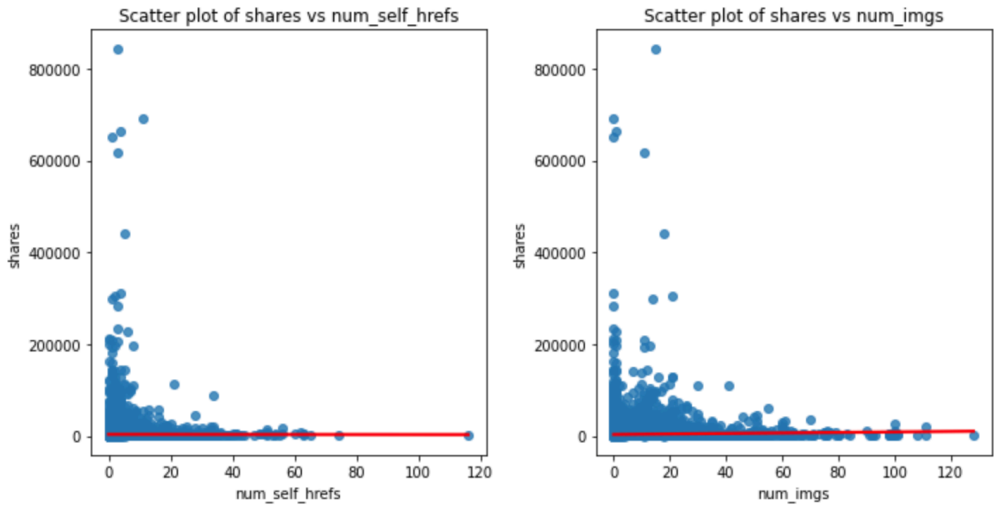
Figure (19) shows the graph of the standardized residuals vs. the dependent variable. The data looks reasonably randomly distributed around the line of y = 0, so the assumption linearity is met.

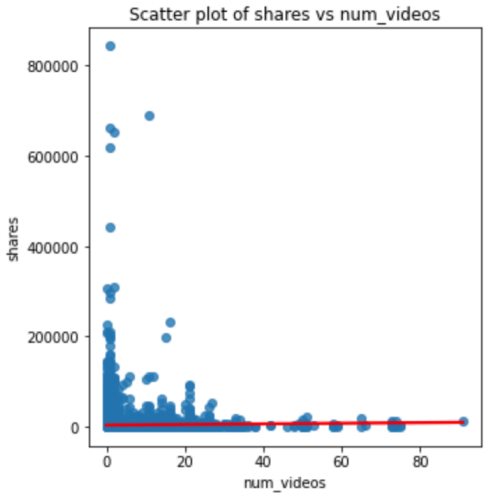
We also ran the Breusche-Pagan test and got a value of $5.77 * 10^{-122}$. So, we reject the null hypothesis of homoscedasticity.

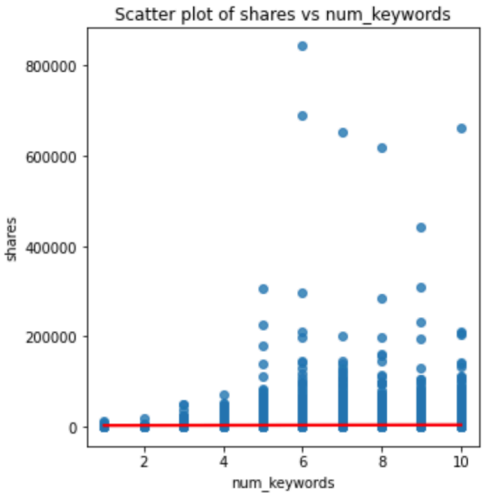
The assumption of linearity did appear to be better satisfied once adding the interaction terms. However, in order to make conclusions about our coefficient terms, we decided to use the log-transformed model without interaction terms.

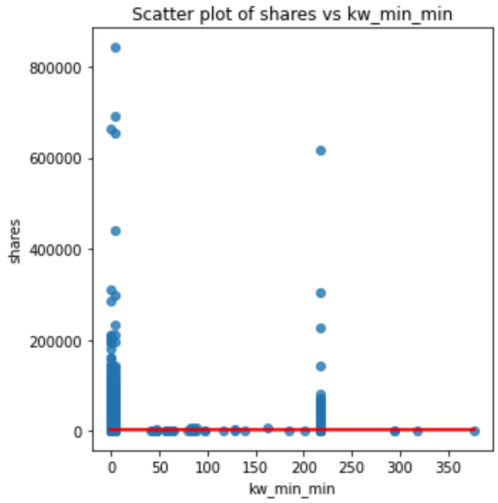
Scatterplots

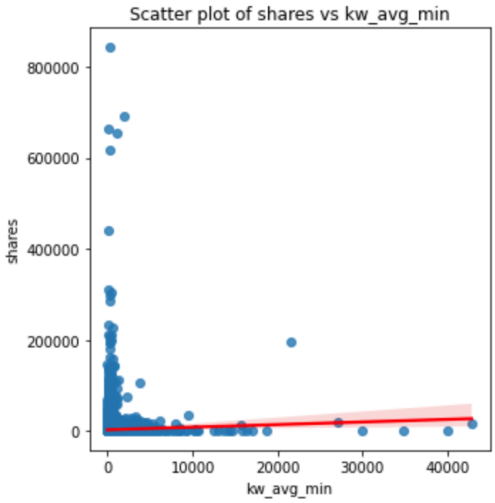
Additionally, we created individual scatterplots graphing each covariate against the dependent variable. This allows us to visualize the relationship between each of the covariates and the dependent variable independently.

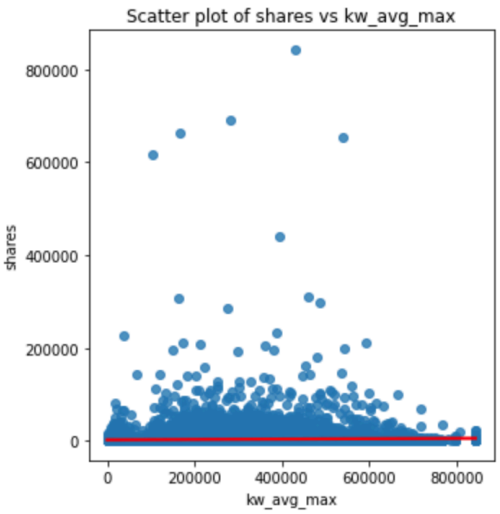


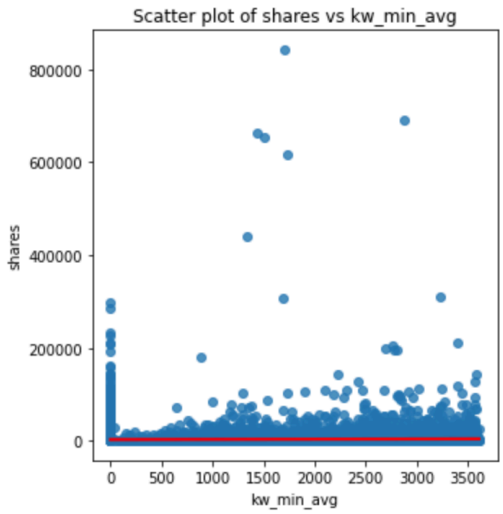

Scatter plot of shares vs num_videos
shares
num_videos
0
200000
400000
600000
800000
0
20
40
60
80

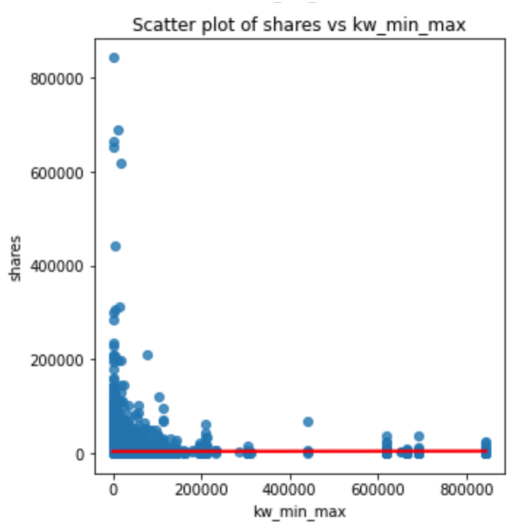

Scatter plot of shares vs num_keywords
shares
num_keywords
0
200000
400000
600000
800000
2
4
6
8
10

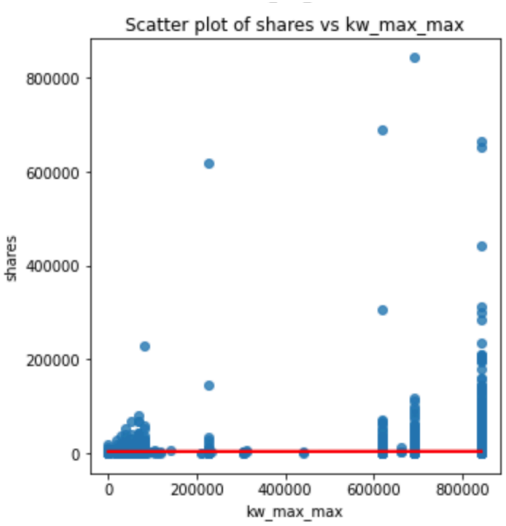

Scatter plot of shares vs kw_min_min
shares
kw_min_min
0
200000
400000
600000
800000
0
50
100
150
200
250
300
350

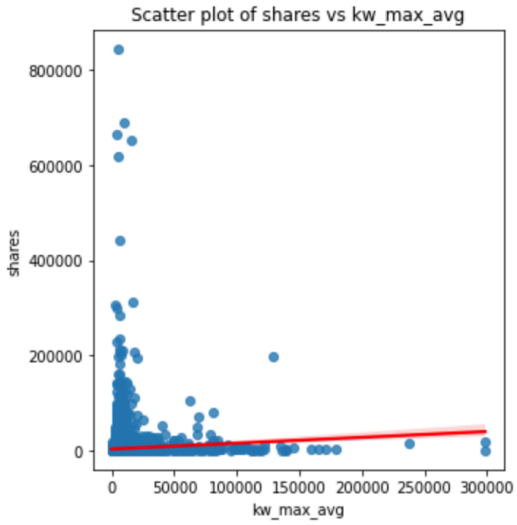

Scatter plot of shares vs kw_avg_min
shares
kw_avg_min
0
200000
400000
600000
800000
0
10000
20000
30000
40000

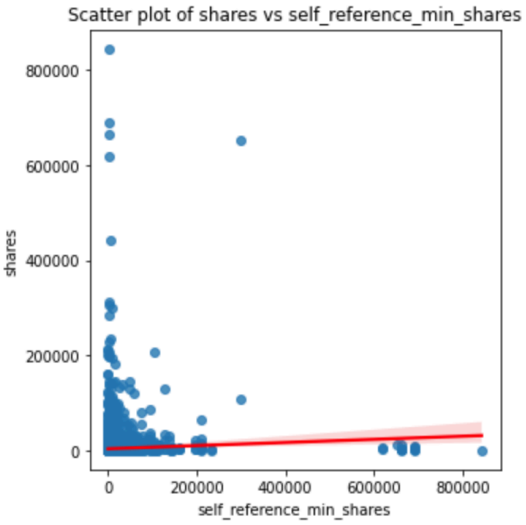

Scatter plot of shares vs kw_avg_max
shares
kw_avg_max
0
200000
400000
600000
800000
0
200000
400000
600000
800000

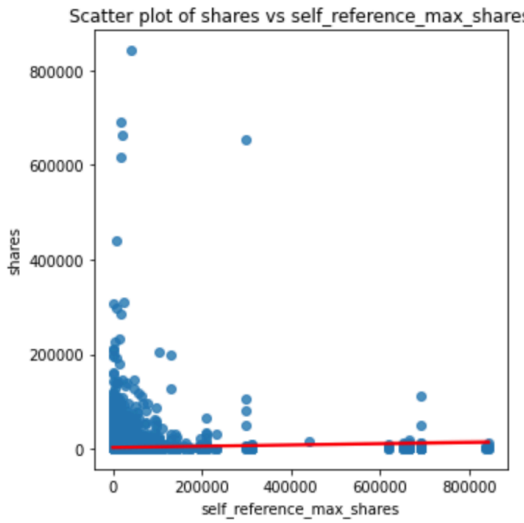

Scatter plot of shares vs kw_min_avg
shares
kw_min_avg
0
200000
400000
600000
800000
0
500
1000
1500
2000
2500
3000
3500


Scatter plot of shares vs kw_min_max
shares
kw_min_max
0
200000
400000
600000
800000


Scatter plot of shares vs kw_max_max
shares
kw_max_max
0
200000
400000
600000
800000


Scatter plot of shares vs kw_max_avg
shares
kw_max_avg
0
50000
100000
150000
200000
250000
300000
400000
600000
800000


Scatter plot of shares vs self_reference_min_shares
shares
self_reference_min_shares
0
200000
400000
600000
800000


Scatter plot of shares vs self_reference_max_shares
shares
self_reference_max_shares
0
200000
400000
600000
800000

These scatterplots show that all of the variables either have a negligible or slightly positive effect on the dependent variable.

Conclusion

Ultimately, we conclude that the variables num_imgs, num_videos, and num_keywords have the greatest effect on the number of shares of an online article. These are the variables that have the greatest coefficient terms while having significant p-values. Holding all other variables constant, adding an additional keyword will, on average, increase the number of shares by a factor of $e^{0.0497} = 1.051$. Additionally, holding all other variables constant, adding an additional video to the article will, on average, increase the number of shares by a factor of $e^{0.0034} = 1.003$. Lastly, holding all other variables constant, adding an additional image to the article will, on average, increase the number of shares by a factor of $e^{0.0081} = 1.008$.

If someone were looking to increase the number of shares on an article they are publishing, we would suggest they consider increasing the number of images, videos, and keywords in their article. In our project proposal, we predicted that the number of images and videos would be positively correlated with the number of shares, and we appear to have been correct. However, we did not predict that the number of keywords would be important to the number of article shares, and it was interesting to challenge our preexisting beliefs during this project.

However, there are several limitations in our models and project. Firstly, all of the data comes from articles published on Mashable. This platform is just one place where online news is published. If we were to look at news published on a different website, there may be different people who visit that website and whose sharing patterns are related to different variables. Thus, the sample might be biased.

Moreover, our assumptions for the regression were not perfect. Although we did logistic transformation on our dependent variable “shares” and added all the two-way interaction terms of the final covariates, some assumptions of the linear regression models were still not met. Normality has improved a lot in the final model compared to the initial model without log transformed dependent variable and interaction terms, but it is still not ideal. According to the Breusch-Pagan test, we still obtained a very small p-value of $9.79 * 10^{-163}$, which led us to reject the null hypothesis of homoscedasticity and implies heteroscedasticity.

Overall, the question of what factors affect online news sharing is one that is of utmost importance in today's digitalized world. Based on our project, we found that the number of images, number of videos, and number of keywords included in the articles are the factors to keep in mind.

References

1. Kari, Dariush et al. “A Novel Family of Boosted Online Regression Algorithms with Strong Theoretical Bounds.” *arXiv: Statistics Theory* (2016): n. Pag. https://www.semanticscholar.org/paper/A-Novel-Family-of-Boosted-Online-Regression-with-Kari-Khan/d87257e5f862f3c033f9a7034864f59721037c8c

2. Ting, Daniel and Eric Brochu. “Optimal Sub-sampling with Influence Functions.” *Neural Information Processing Systems* (2017). https://www.semanticscholar.org/paper/Optimal-Sub-sampling-with-Influence-Functions-Ting-Brochu/aa500cfa29afa3ba5a6440855575314d1e2b6b8a