Predicting the popularity of internet posts

Athens University of Economics and Business

Master of Science in Business Analytics

Statistics for BA I

Christos Vlassis

f2822204

Contents

1. Abstract

Understanding the key factors that drive an article to each potential is a topic that many researchers have investigated. The main goal of this paper is to identify the factors that make a post viral and create a regression model to predict the shares of other articles with the use of their characteristics. In total seven models were created all showing small Adjusted R Squared and poor RMSE scores. The model that was chosen showed small violation of the regression assumption and 10-fold cross validation was used to test its predictive capability.

2. Introduction

Regression models were first introduced in 1805 by Legendre with the least square method. Their aim was to determine the bodies that orbit the sun. Later on, in history, it was used for solving biological problems such as how height of parents affects the height of their offsprings. In recent years the regression models are often used in the field of economics. For example, the gravity model, a well-known model for interpreting the bilateral trade balance of two or more countries. In the case of this study a regression model will be used for predicting the shares of a post. Also, we will try to find what makes a post viral and the significance of its characteristics to its shares.

3. Literature review

A current interest in predicting the popularity of web material has attracted many researchers to make predictions and try to understand what are those features that drive an article to its potential. The study of [1] predicts the news popularity before any publication was made by analysing five main features. Length characteristics of the article, time, authors and category scores ad finally features such as top people name and verbs (this are some of the data that they used). They found that the most important features for predicting the article polarity is the author something that shows that the authors writing has an effect in the article popularity.

Other studies such as the the work of [2] proposed an IDSS that extracts a set of features of an article and predict the future popularity. They used characteristic same as the dataset that is analysed by this paper and they had the best score using Random Forest. they found that keyword-based features Natural Language Processing features and previous shares are the most important features of a good post.

In a study employing real data from the UCI Machine Learning Repository, the author employed LDA to lower the dimension as well as three different learning algorithms, including AdaBoost, LPBoost and Random Forest, to find which is the best model to predict the popularity of online news. The Adaptive Boosting model produced the greatest results.[3].

A study by Choudhary sought to increase the rate of popularity forecast of an article. Using a dataset of approximately 40000 articles, 60 characteristics, and one class label attribute from the UCI Machine Learning Repository . To find the ideal collection of characteristics that ought to be taken into account when writing an essay, an algorithm was applied. Naive Bayes

had the greatest prediction with an accuracy of 93%, while neural networks had the highest prediction value for 18 attribute sets with an accuracy of 91% [4].

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Serge Fdida [5] addressed the problem of predicting an articles popularity based on the comments made by the users. They analysed the ranking performance of three prediction models by using data from articles that covered a four year period. Their results showed that prediction methods improve the ranking performance and observed that a simple linear regression prediction method outperformed more dedicated prediction methods.

4. Methods, data acquisition & Exploration

4.1 Data acquisition and exploration

The data were given to us by Mr. Ntzoufras and contain 61 Attributes for different articles that were published in Mashable. The data are splitted into two datasets, the training dataset that is used for the training of the regression model and the test dataset that is used for testing the predictive capability of the model. Also, the training dataset contains 3000 observations and the test dataset contains 10000 observations. Two columns were removed from both datasets, the URL of the article and the 'timedelta' which refer to the days between the article publication and the dataset acquisition. These columns were removed because they don't have any predictive capability in the model.

The data can be described in 7 categories each with its own characteristics. The following table explains the categories of the data and their types. Some of the types are integers some ratios and in other cases they can be considered as dummy/categorical variables.

Table 1: Data categories and data types

Attribute	Data type	Category	Data type
Article type		**Shares and links**	
Is Lifestyle/ Tech Entertainment/ World Business/ Media	Categorical	Min/Max/Avg shares of referenced article in Mashable	integer
Videos and Images		Num of links	integer
Number of images	Integer	Num of links to other articled in	integer
Number of Videos	integer	**Day of published article**	
Key Words		Day published	Categorical
		Is published on weekend	Dummy
Num of keywords	integer	**Words**	
Worst keyword (max,min avg Shares)	integer	Num of words in title	Integer

Best keyword (max,min avg Shares)	integer	Num of words in content	Integer
Avg keyword (max,min avg Shares)	integer	Rate of words in content	ratio
Avg length of words	integer	Rate of non-stop words in content	integer
Natural Language Processing			
Avg/Min/Max Positive Polarity	Ratio		
Avg/Min/Max Negative Polarity	Ratio		
LDA	Ratio		
Text subjectivity	Ratio		
Sentiment polarity	Ratio		
Global Positive/negative words rates	Ratio		
Positive/negative words rates	Ratio		
Title polarity/subjectivity	Ratio		
Absolute polarity/subjectivity level	Ratio		

As it can be seen from Table 1 the dataset contains 58 variables and the variable shares which is the dependent variable of the regression that is about to be made. Due to the large number of attributes creating a correlation table on this point is unfeasible.

In table 2, on the appendixes, we can see that there is small correlation between the dependent variable ‘shares’ with all the other continuous variables.

To have a better understanding of the categorical variables two Graphs have been created. The first Graph shows the number of observations for each of the channel type and the second Graph show the number of articles that were published for each day of the week.

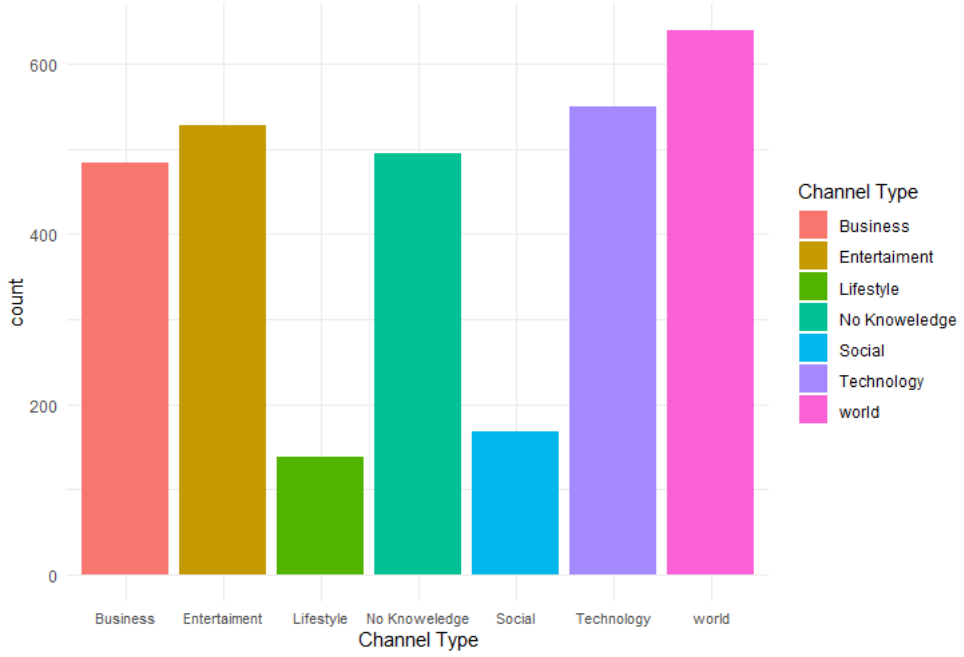
4.1.1 Categorical variables

In the dataset two main categories can be found. The first category contains the day that the article was published. The second main category contains the channel information. In this section the categorical variables will be analysed in order to have a better understanding of how the categories affect the articles shares.

Graph 1 shows the number of URLs per channel type (we use the training dataset). It seems that the majority of the articles concern world news. Technology follows with 580

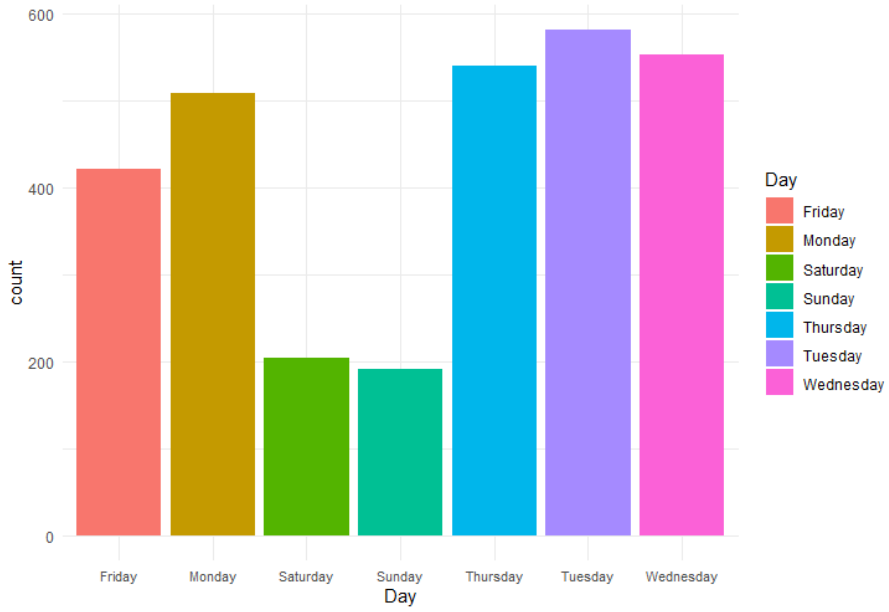
articles and Lifestyle news have the least number of observations. It is worth mentioning that 550 articles seem to not belong to any of the categories, meaning that there is no knowledge of their thematic area.

Graph 1: Number of articles per Channel Types



Graph 2 shows the number of articles that were published for each day of the week. Most of the articles were published on Tuesday followed by Wednesday and Thursday. On Saturdays and Sunday approximately, only 410 articles were published.

Graph 2: Number of articles per day



Graph 3 (in the appendix) contains the box plots for the shares variable for each channel category. Due to the large range of values of the shares variable (minimum value 8, maximum value 617900) the box plots can't give us any useful information.

In order to have a better understanding of the relationship between the Shares of each category of the articles the use of pairwise comparison was made. The pairwise tests helps us understand this relationship, with the null hypothesis being that there is no difference between the medians. The statistical evidence shows that there is no difference between the medians of the shares for the following combination of channels. Lifestyle and Business, 'No Channel Knowledge' and Lifestyle, Social and 'No Channel Knowledge', Social and 'No Channel Knowledge', Technology and Lifestyle, Technology and 'No Channel Knowledge' and finally World and Entertainment. For the other combination of channels there is statistical evidence that shows that there is significant difference between the medians of the shares. Also, the 'World' channel type seems to contain the lowest number of shares, on average. Finally, on average, the majority of the shares seems not to contain information about the Channel Type.

The dataset, also, provides us the information of which day of the week the article was published. With that information a comparison was be made for the medians of each day of the week for the depended variable shares. All the relevant information can be found in Table 4 in the appendix. One of the main things that can be noticed is that there is no statistically significant difference in the medians of the shares for the weekend. Finally, on average, the majority of shares are made on Saturday, with Wednesday having the least number of shares.

4.1.2 Continuous and Discrete variables

Analysing the distribution of the continuous and discrete variables gives us a better understanding of data that are in possession. In this subsection a brief discussion is going to be made about the histograms and correlation of the continuous and discrete variables. Due to the large number of continuous and discrete variables, 21 in total, all the graphs and tables that are used can be found in the appendix section.

The depended variable shares isn't normally distributed. This fact hints us, that in the modelling section of this paper the use of logarithms should be made to solve some of the regression problems that might occur. If the regression assumptions are not met, then wrong conclusions will be made for the interpretability of the model. The rest of the continuous variables don't show any signs of normality with exemption of the 'n_tokens_title'.

Understanding the relationship between the dependent variable 'shares' with all the other continuous variables is crucial for every regression model. The use of scatter plots will give all the relevant information as seen in Graph 6 and Graph 7. As it can be seen, no direct relationship can be found between the dependent and all the other independent variables. This proves that the correlations are very low. To resolve this issue the use of logarithms may be proven useful in the modelling section since the shares variable has a very large range of values in comparison with the other independent continuous variables.

Concluding this subsection, the following can be said: the raw data, the data given to us with no transformations, show no normality in most cases. Also, the correlation between the dependent and independent variables is extremely low, indicating that there is no relationship. Finally, approximately 500 articles do not belong to any channel type, meaning that there is no information about the content of the article. With that being said, the use of logarithms might help to resolve many of these issues and result in a better model.

5. Modelling & Methodology

5.1 Methodology & Modelling comparison

The main aim of this paper is the creation of a regression model for predicting the popularity of an article depending on some key factors. These factors can be found in Chapter 4. The secondary purpose of this paper is to create a model capable of interpretating the number of shares based on the independent variables. With that in mind, a model was created that had as its main priority the prediction and as its secondary the interpretability.

In total, six models were created and tested all with some key differences. The models with their descriptions can be found in the appendix. In all models the AIC method was used since it is recommended for prediction modelling. Also, the stepwise procedure was used as it gives the best results. In three out of six models a Lasso Regression was made in order to find the best variables for the min lambda and later the AIC was implemented. In the other three models, only the AIC was used for the selection of the variables. A 10-fold cross-validation was used to find the best prediction model. Also, the regression assumptions were checked in order to have a better understanding of the interpretability of the models. Since the main goal is finding the best model for prediction, we focused on the model that had the best 10-fold cross-validation score (model3). Finally, in all models the Residual standard error was approximately 0.85 and the Adj R squared 0.13.

To have a better understanding of our model the cook distance was found (mean = 0.000357). From Graph 11, in the appendix, we find the outliers of our data that was used in the model. We found that 227 observations are being considered as influence points, using 3x the mean of the cook's distance as a cut of point.

Table 7: First look at the results of the models

Models:	**Methods**	**Data Changes**	**Adj-R Squared**	**Residual Standard Error**	**RMSE of 10-Fold Validation**
Model 1	AIC	logged shares	0.1316	0.8505	0.8564
Model 2	AIC	No changes	0.1004	0.8657	10953.69
Model 4	Lasso, AIC	No changes	0.1295	0.8515	10841.29
Model 5	Lasso, AIC	logged shares	0.1316	0.8505	0.8561

Model 6	Lasso, AIC	Logged all continuous variables	0.1471	0.8429	0.8457
Model 3	AIC	Logged all continuous variables	0.1501	0.8414	0.8446

From Table 7, model 3 shows the most potential. With that in mind, the rest of the analysis was focused, mostly, on Model 3. Model 2 and Model 4 had many issues regarding the linearity assumptions. Model 1 and Model 5 had similar results but also had issues regarding the linearity assumptions. Furthermore, all models except model 3 had violations in most of the regression assumptions. Finally Model 6 and Model 3 had the best results, with Model 3 being the best in comparison to all others.

5.2 The Main Model

5.2.1 The Model

Model 3 was chosen as the main model for this paper. Firstly, there were some issues regarding multicollinearity. The variables 'self_reference_max_shares' and 'self_reference_avg_sharess' had 0.95 correlation so 'self_reference_max_shares' was dropped. Also, the following variables were removed because they were not statistically significant, 'data_channel_is_lifestyle' and'data_channel_is_bus'. Finally, by logging the continuous variables a new issue arrived, N/As and Inf values were found. To cope with this issue the mean of each variable was used to replace these values. With these changes the following model was created:

Model:

Log(Shares) =1.4498 - 0.155 * log(n_tokens_content) -0.9355 * log(n_unique_tokens) + 0.6190 * log(n_non_stop_words)+ 0.0567 * log(num_imgs) + 0.1729 * log(num_keywords) - 0.2392 * log(data_channel_is_entertaiment) + 0.2383 * log(data_channel_is_socmed) - 0.1536 * log(data_channel_is_world) + 0.0820 * log(kw_min_min) + 0.4149 * log(kw_min_avg) + 0.3885 * log(kw_avg_avg) + 0.1348 * log(self_reference_avg_sharess) - 0.2385 * log(weekday_is_monday) -0.2753 * log(weekday_is_tuesday) -0.3285 * log(weekday_is_wednesday) - 0.2756 * log(weekday_is_thursday) - 0.2054 * log(weekday_is_friday) - 4.0372 * log(global_rate_negative_words) - 0.5688 * log(min_positive_polarity) - 0.2727 * log(avg_negative_polarity) + ε

$$\varepsilon \sim N(0, 0{,}8438^2)$$

Table 10: Significant tests for independent variables

Variable	Statistical significant
n_tokens_content	Statistical significant at a=0.01
n_unique_tokens	Statistical significant at a=0.01
n_non_stop_words	Statistical significant at a=0.01
num_imgs	Statistical significant at a=0.05
num_keywords	Statistical significant at a=0.01
data_channel_is_entertaiment	Statistical significant at all levels of a

data_channel_is_socmed	Statistical significant at all levels of a
data_channel_is_world	Statistical significant at all levels of a
kw_min_min	Statistical significant at all levels of a
kw_min_max	Statistical significant at a=0.01
kw_min_avg	Statistical significant at all levels of a
kw_avg_avg	Statistical significant at all levels of a
self_reference_avg_sharess	Statistical significant at all levels of a
weekday_is_monday	Statistical significant at all levels of a
weekday_is_tuesday	Statistical significant at all levels of a
weekday_is_wednesday	Statistical significant at all levels of a
weekday_is_thursday	Statistical significant at all levels of a
weekday_is_friday	Statistical significant at all levels of a
global_rate_negative_words	Statistical significant at a=0.05
min_positive_polarity	Statistical significant at a=0.05
avg_negative_polarity	Statistical significant at a=0.05
(Intercept)	Statistical significant at a=0.05

#Note. Interpretation of the significance tests: With a P-value <0.05 we reject the null hypothesis. So there is statistically significant evidence that the changes in the independent variables affect the dependent variable (shares).

The final model was based on model 3 since it had the most promising results. After the removed variables. the new model had the following characteristics. The Residuals standard error changed to 0.8438 and the Adjusted R-Squared to 0.1452 which shows how well the independent variables explain the variability of the dependent variable shares. The 10-fold cross-validation shows a Root mean square of 0.8457. As it can be seen the new and final model has worst scores, but no multicollinearity is present in it. Finally, from Graph 9, in the appendix, it can be seen that with the use of logarithms the correlation between shares and the rest continuous variables has increased.

5.2.2 Assumption Control

As seen in the previous subsection the **multicollinearity** assumption is not violated. Some small indications of multicollinearity can be seen only for the variables 'n_tokens_content' (VIF = 5.3) and 'n_unique_tokens' (VIF = 8.9). But, in either case they don't pass the value of 10, so both variables will be kept in the model.

From the Residuals vs Fitted plot of Graph 8, it can be seen that the **linearity** assumption is mostly met. Some departure can be found in the right part of the tail. By removing some of the extreme observations, such as those that are suggested by R(777,2290,2173), a more linear model would be created. For the Residual vs Leverage Graph, a linear line can be seen with some extreme values such as 2290,777 and 721.

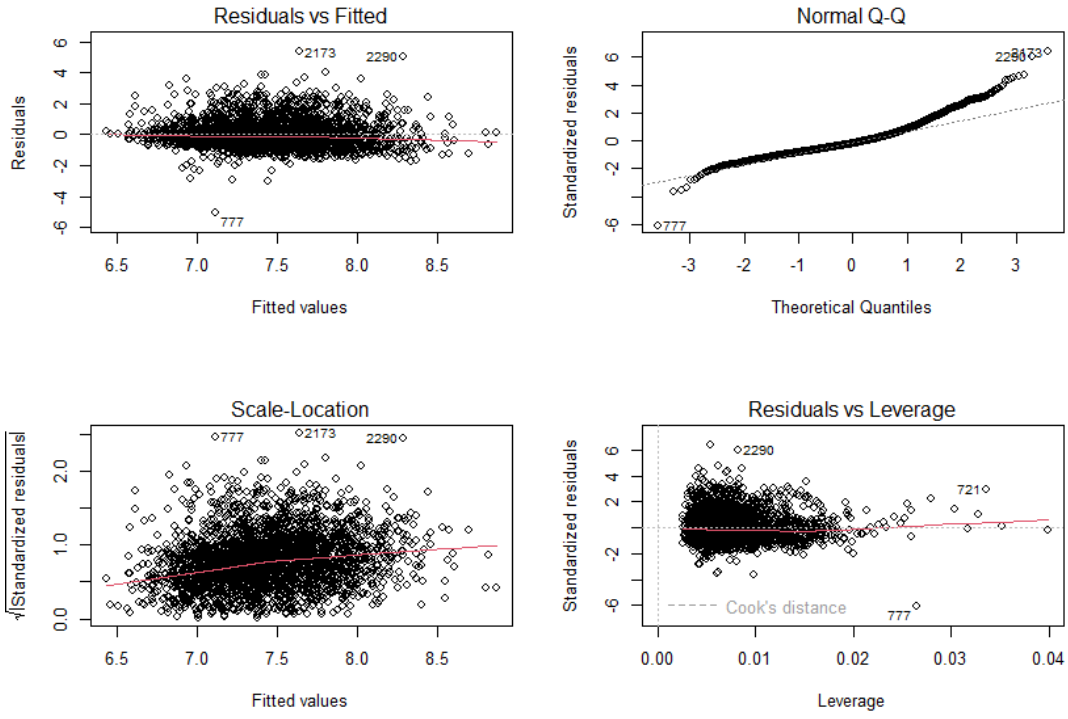
The **normality** assumption, as seen from the Normal Q-Q is not fully met. A departure from it can be seen in the right tail of the line. More specifically by making use of Kolmogorov-Smirnov and Shapiro-Wilk test for the normality of the residuals the p-Value (2.2e-16) is less than 0.05. With that being said, we reject the null hypothesis. We have statistical significance

evidence to show that the residuals do not follow normality. But, due to the fact, that there is a large sample QQ plot will give us more trustworthy results than the tests.

The **homoskedasticity** assumption is not met. The variance of the residuals is not constant and is large. The Levene test, for the equality of the variance, had a p-value equal to 8.578e-16, less than 0.05. So, we reject the null hypothesis meaning that we have statistically significant evidence to show that the variance of the model is not constant. Also, as can be seen from Graph 10 in the appendix the variance is large and not constant.

Overall, the model follows linearity and based on the rule of thump that VIF should not exceeded 10 there are no multicollinearity issues. But, with the normality of residuals not fully met the p-values might be misleading. Finally, the variance of the residuals, as it can be seen from the Residual vs Fitted Graph an from Graph 10 is not constant and large meaning that the homoskedasticity assumption is not met.

Graph 8: Model Plots



5.2.3 Predicting using the Test Data

To test our model, the use of the test data set was made to find the MSE. The same data changes that were made in the training data set were made in the test data set. The MSE score is not satisfying, MSE = 0.76. Also, the predicted R squared is 0.14 indicating another time the poor predicting capability of the model. These scores show the predictive capability of our model, which is poor.

6. Analysis and result interpretation

The raw data had many issues regarding the regression assumptions, so, many changes were made to have a well fitted model. The summary of the model can be found in the appendixes, Table 9. This section contains the coefficient interpretation. Due to the large number of variables not all variables will be interpreted but the tables 11 & 12 contain the coefficients values for further investigation.

6.1 Continuous variables

Due to the use of logarithms in all variables, dependent and independent, the interpretation of the coefficients has changed. For the variable n_tokens_content, a 1% increase in the number of tokens in the content will a decrease the number of shares by 0.155%, on average, holding all other variables constant. Also, for the variable n_unique_tokens, a 1% increase in the number of unique tokens in the content will a decrease the number of shares by 0.9355%, on average, holding all other variables constant. For the variable num_imgs, a 1% increase in the number of images will a increase the number of shares by 0.0567%, on average, holding all other variables constant. The same logic can be applied to all other continuous independent variables, all having the same interpretation structure. What follows is the independent continuous variables with their coefficients.

Table 11: Independent continuous variables and coefficients, in ascending order

Variables	Coefficients
global_rate_negative_words	-4.0372
n_unique_tokens	-0.9355
min_positive_polarity	-0.5688
avg_negative_polarity	-0.2727
n_tokens_content	-0.1559
kw_min_max	-0.0820
num_imgs	0.0567
kw_min_min	0.0820
self_reference_avg_sharess	0.1348
num_keywords	0.1729
kw_avg_avg	0.3885
kw_min_avg	0.4149
n_non_stop_words	0.6190
Intercept	1.4498

6.2 Categorical/Dummy Variables

The categorical independent variables have different interpretations than the continuous variables. On average, holding all other variables constant an article that was published on Monday will command 0.238% less shares than an article that was not

published on Monday. Also, on average, holding all other variables constant an article that is from the channel entertainment will command 0.2392% less shares than an article that was not from channel entertainment. The same logic can be applied to all other categorical independent variables, all having the same interpretation structure.

Table 12: Independent categorical variables and coefficients, in ascending order

Variables	Coefficients
weekday_is_wednesday1	-0.3285
weekday_is_thursday1	-0.2756
weekday_is_tuesday1	-0.2753
data_channel_is_entertainment	-0.2392
weekday_is_monday1	-0.2385
weekday_is_friday1	-0.2054
data_channel_is_world	-0.1536
data_channel_is_socmed	0.2383

7.Conclusions

The main goal of this assignment was the creation of a regression model to predict the shares of an article. The secondary goal of this assignment was the creation of a model which has strong interpretability, meaning that the regression assumptions must mostly be met. Finally, a key question must be answered, 'What makes a post viral based on its characteristics?'.

The most important factor, according to the model, is the rating of non-stop words in the content. Also, the best keywords and the average keywords of the article have a positive impact on the number of shares of the article. Social media channels, the number of keywords, the average shares of the referenced article in Mashable and the number of images have positive effects in the popularity of the articles. With that being said, the greater the values of the above variables the greater the expected popularity of the articles.

On the other hand, according to the model, an article losses popularity if it has a high rating in negative words, a high rating of unique words in the content, a high average polarity of negative words and many words in the content. Also, articles that are published on a weekday have less shares than those that are published on the weekend. Finally, the channel type Entertainment and World have a negative impact on the number of shares. In conclusion, these factors should be minimized, and the articles should be uploaded on the weekend.

The predictive capability of the model is not sufficient so other methodologies should be used for prediction in this field and the R squared is low. Only 15% of the variability of shares is explained by our model. Also, the regression assumptions, overall, have not been met, meaning that the interpretability of the model and its p-values cannot be trusted for the most

part. Other researchers had made different models using techniques such as Random Forest which had better predicting results such as [6]. Techniques like these are more advanced beyond our scope of analysis. With that being said, further research is recommended with different techniques to get more robust results.

8. Appendix

Table 2: shows the correlation between the 21 continuous variables before any model was applied.

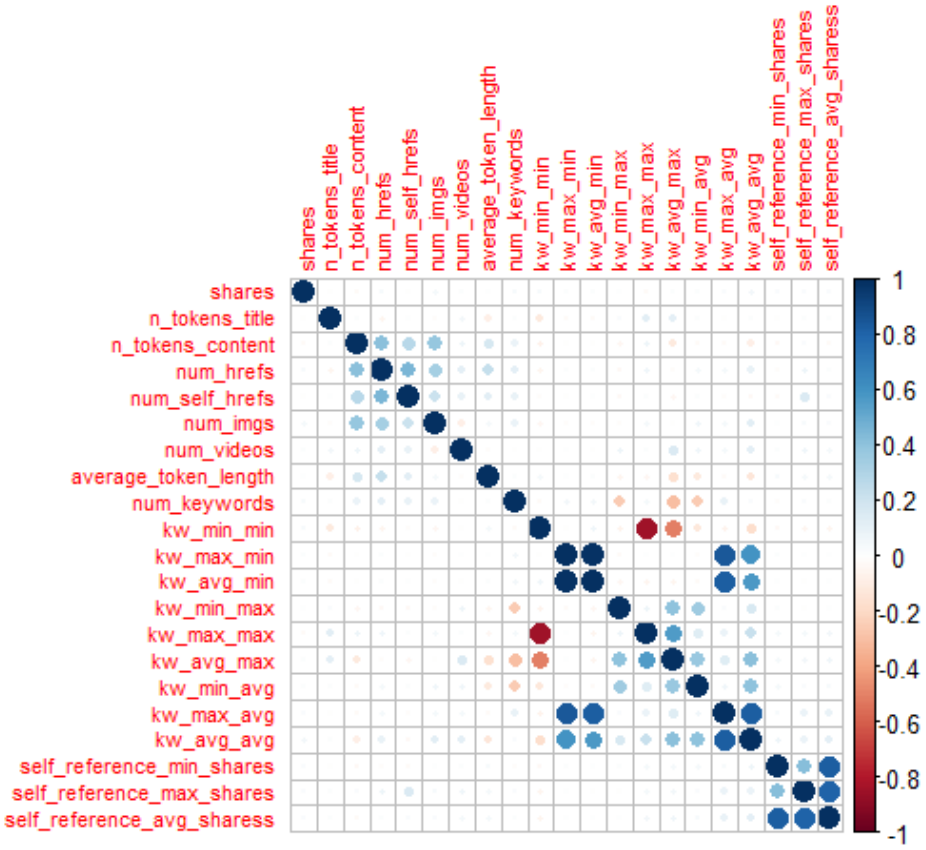


Table 3: shows the pairwise comparison using Wilcoxon rank, that was made between the dependent variable (Shares) with the Channel categories.

 # Note: Normality was rejected, the sample was large and the mean was not a sufficient a sufficient descriptive measure of central location for all groups. With that being said we used the Wilcon pairwise test. When p-Value is less than level of significant (e.g. a=5%) then we reject the null hypothesis. This means that, there are statistically significant evidence that show that there is a difference in the medians.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  Main_Training_for_bar_plot2222$shares and Main_Training_for_bar_plot2222$Type

              Business Entertaiment Lifestyle No Knoweledge Social  Technology
Entertaiment  0.00038  -            -         -             -       -
Lifestyle     0.29682  0.00115      -         -             -       -
No Knoweledge 5.2e-08  < 2e-16      0.13638   -             -       -
Social        1.8e-10  < 2e-16      0.00514   0.29682       -       -
Technology    1.9e-05  < 2e-16      0.48776   0.29682       0.00172 -
world         0.00027  0.62050      0.00037   < 2e-16       < 2e-16 < 2e-16

P value adjustment method: holm
```

Graph 3: contains the boxplots for shares and each category of channel. As can be seen no valuable information is given by the box plots.

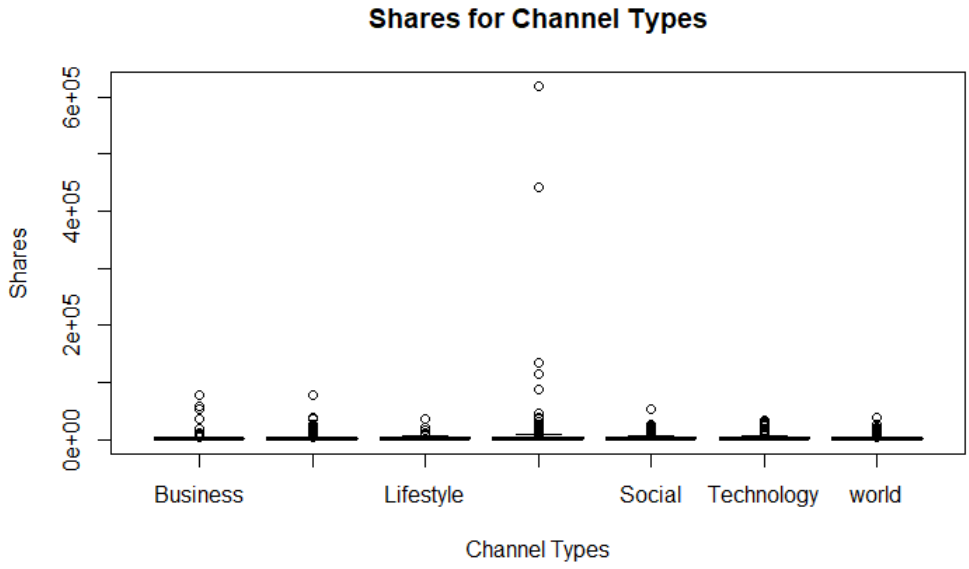


Table 4: shows the pairwise comparison using Wilcoxon rank, that was made between the dependent variable (Shares) for each day of the week.

 # Note: Normality was rejected, the sample was large and the mean was not a sufficient a sufficient descriptive measure of central location for all groups. With that being said we used the Wilcon pairwise test. When p-Value is less than the level of significant (e.g. a=5%) then we reject the null hypothesis. So there are statistical significant evidence that show that there is a difference in the medians.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  Main_Training_for_bar_plot_day$shares and Main_Training_for_bar_plot_day$Day

          Friday  Monday  Saturday Sunday  Thursday Tuesday
Monday    1.00000 -       -        -       -        -
Saturday  2.4e-06 1.0e-07 -        -       -        -
Sunday    0.00083 3.8e-05 1.00000  -       -        -
Thursday  1.00000 1.00000 3.3e-09  3.7e-06 -        -
Tuesday   0.40553 1.00000 6.4e-10  7.9e-07 1.00000  -
Wednesday 0.21635 0.95315 2.2e-11  8.3e-08 1.00000  1.00000
```

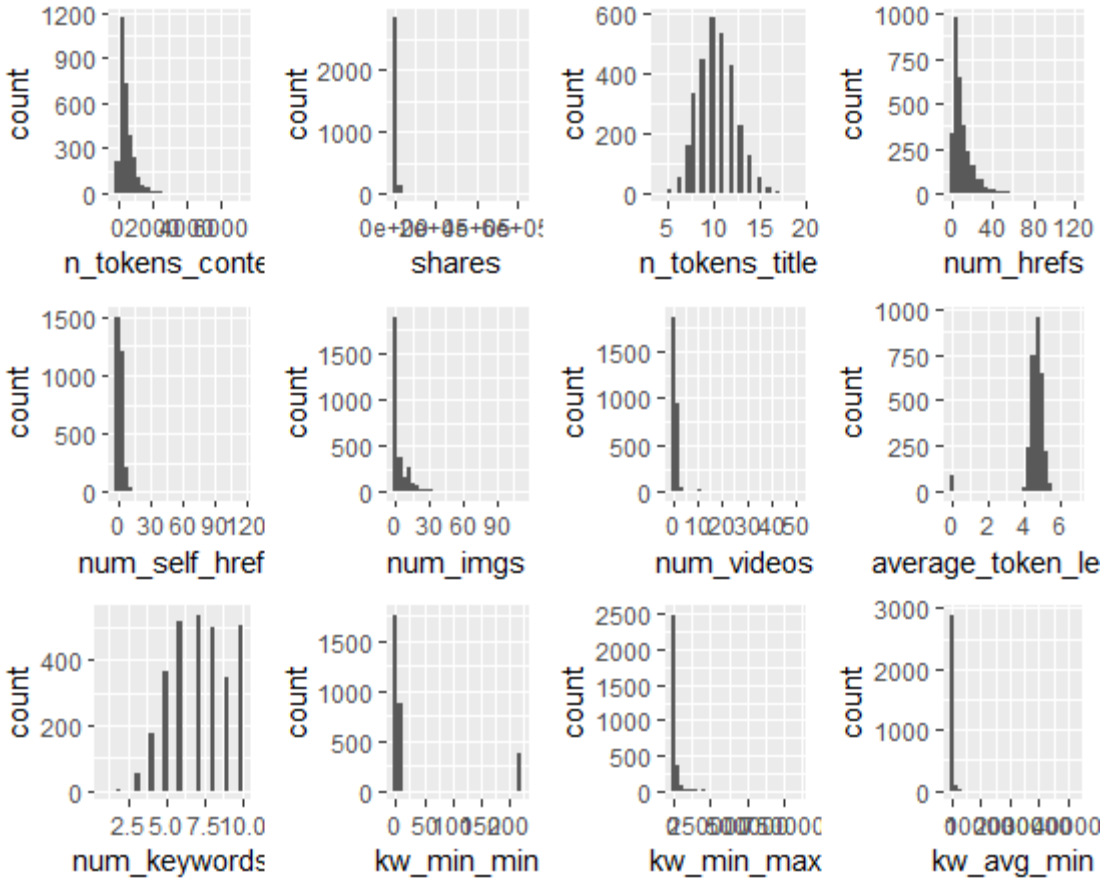
Table 5: Mean Shares per Channel type

	Group.1	x
7	world	1854.756
1	Business	2532.153
2	Entertaiment	2667.598
6	Technology	2924.758
3	Lifestyle	2959.101
5	Social	4207.780
4	No Knoweledge	6921.257

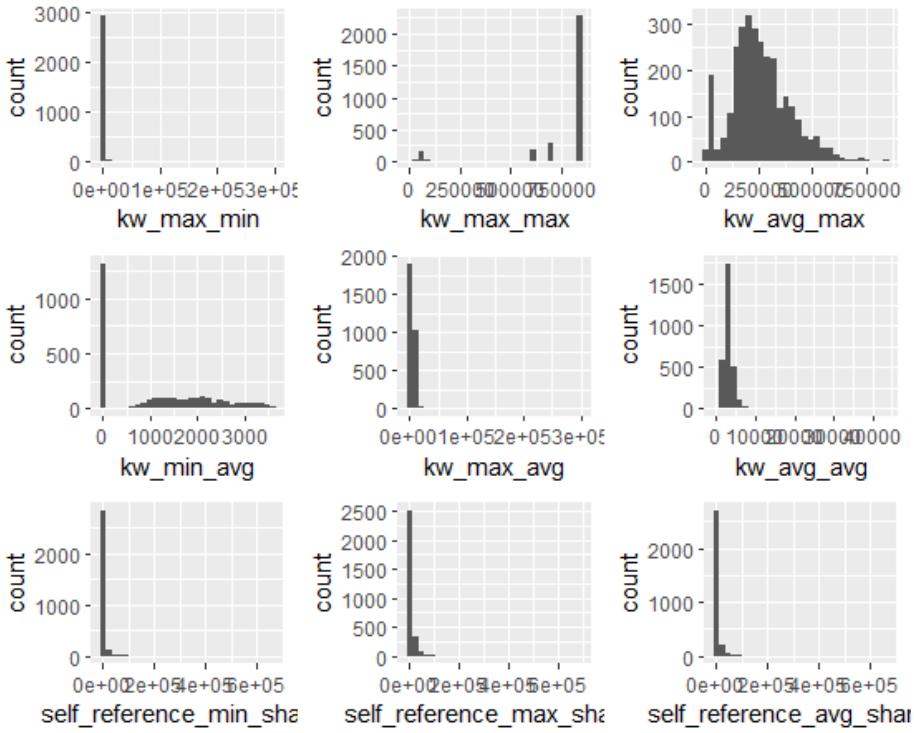
Table 6: Mean Shares per day

	Group.1	x
7	Wednesday	2625.844
5	Thursday	2904.656
2	Monday	2955.193
1	Friday	3089.147
4	Sunday	3320.516
6	Tuesday	3745.224
3	Saturday	6473.902

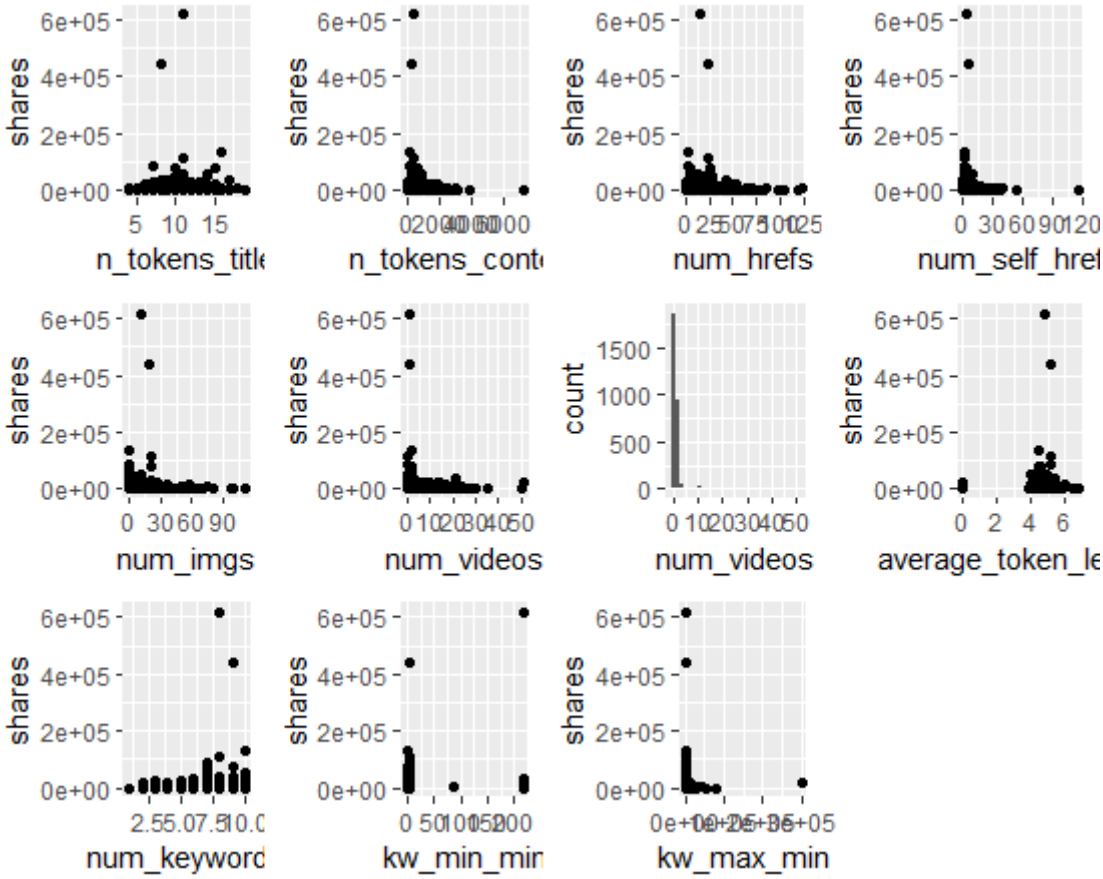
Graph 4: Show the distribution of the first 12 continuous variables



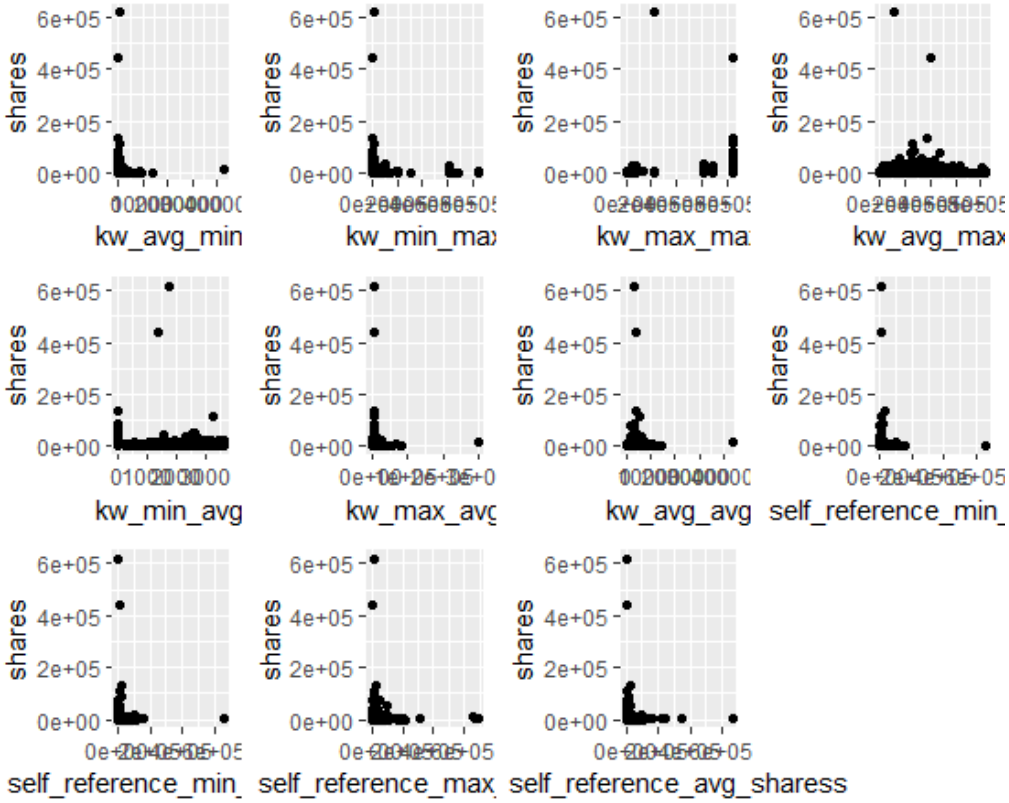
Graph 5: Show the distribution of the 9 continuous variables



Graph 6: Shows the Scatter plots between the Shares with 11 variables



Graph 7: Shows the Scatter plots between the Shares with 11 variables



Models:

1. **method AIC with direction both.We also use the log of shares** MODEL1 # Model 1:Residual standard error 0.8505, Adj R Squared 0.1316

   RMSE = 0.8564537

2. **method AIC with direction both. No changes were made in the data**

Model 2: Residual standard error: 0.8657, Adjusted R-squared: 0.1004

RMSE = 10953.69

3. **AIC method, direction both. All continuous variables were logged and N/As were replaced with columns means.**

MODEL 3: Residual standard error: 0.8414, Adjusted R-squared: 0.1501

RMSE = 0.8446748

Main Model: Residual standard error: 0.8438, Adjusted R-squared: 0.1452

RMSE = 0.8472157

'Main Model' is the model that was finally used for the analysis. This model was created by removing the variables that were not statistically significant and the variables that was causing multicollinearity. It is the final model.

4. **1st we use lasso regression, then AIC method with direction both, to the raw data.**

Model 4: Residual standard error: 0.8515, Adjusted R-squared: 0.1295

RMSE = 10841.29

5. **1st we used lasso regression, then AIC method with direction both. We used log of shares.**

Model 6: Residual standard error: 0.8505, Adjusted R-squared: 0.1316

RMSE = 0.8561833

6. **1st we used lasso regression, then AIC method with direction both. All continuous variables were logged and N/As where replaced with columns means**

Model 6: Residual standard error: 0.8429, Adjusted R-squared: 0.1471

RMSE = 0.8457192

Table 8: VIF scores of the Final Model

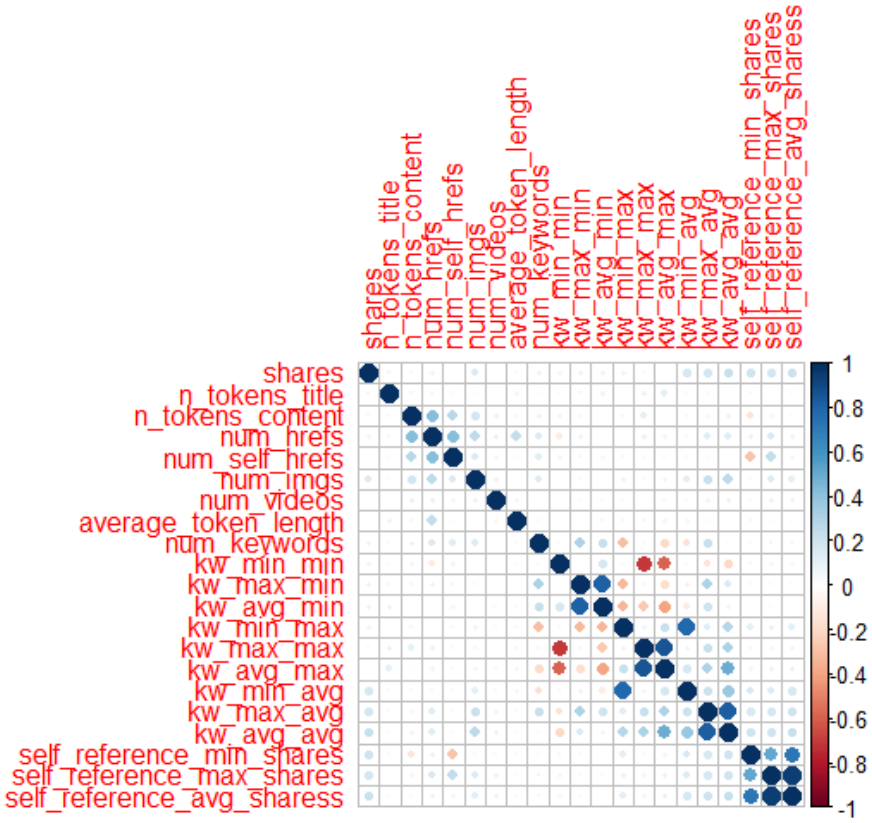
```
> vif(model3_TRY)
          n_tokens_content              n_unique_tokens              n_non_stop_words                    num_imgs
                  5.314173                     8.980467                     5.250243                    1.247817
             num_keywords data_channel_is_entertainment       data_channel_is_socmed        data_channel_is_world
                  1.194202                     1.142457                     1.082041                    1.298129
               kw_min_min                   kw_min_max                   kw_min_avg                   kw_avg_avg
                  1.064621                     2.882574                     2.947221                    1.472403
 self_reference_avg_sharess           weekday_is_monday            weekday_is_tuesday          weekday_is_wednesday
                  1.110695                     1.928503                     2.007423                    1.987060
       weekday_is_thursday             weekday_is_friday    global_rate_negative_words        min_positive_polarity
                  1.967023                     1.800158                     1.198908                    1.360416
     avg_negative_polarity
                  1.261429
>
```

Table 9: Summary of Model

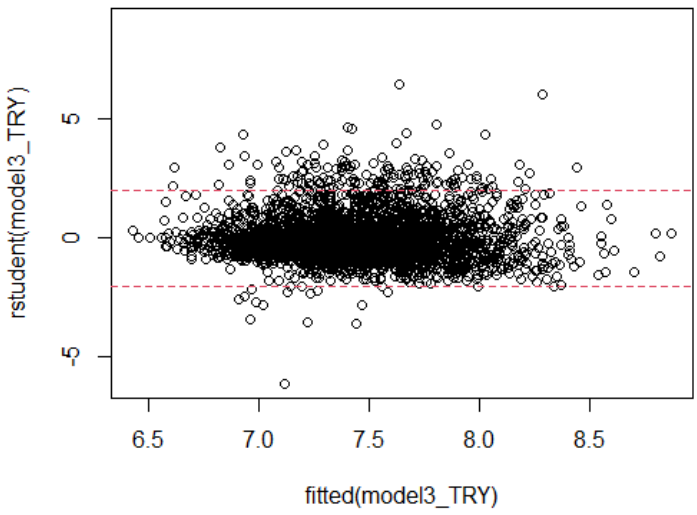
```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.44984    0.64092   2.262 0.023761 *
n_tokens_content               -0.15598    0.04991  -3.125 0.001795 **
n_unique_tokens                -0.93550    0.33490  -2.793 0.005249 **
n_non_stop_words                0.61902    0.20920   2.959 0.003111 **
num_imgs                        0.05679    0.01600   3.550 0.000392 ***
num_keywords                    0.17296    0.05839   2.962 0.003080 **
data_channel_is_entertainment1 -0.23922    0.04327  -5.528 3.51e-08 ***
data_channel_is_socmed1         0.23832    0.06970   3.419 0.000636 ***
data_channel_is_world1         -0.15360    0.04287  -3.583 0.000345 ***
kw_min_min                      0.08203    0.01355   6.052 1.61e-09 ***
kw_min_max                     -0.08203    0.02539  -3.231 0.001248 **
kw_min_avg                      0.41493    0.07928   5.234 1.77e-07 ***
kw_avg_avg                      0.38859    0.05507   7.056 2.12e-12 ***
self_reference_avg_sharess      0.13483    0.01688   7.990 1.92e-15 ***
weekday_is_monday1             -0.23859    0.05700  -4.186 2.92e-05 ***
weekday_is_tuesday1            -0.27538    0.05524  -4.985 6.54e-07 ***
weekday_is_wednesday1          -0.32857    0.05605  -5.863 5.06e-09 ***
weekday_is_thursday1           -0.27564    0.05624  -4.901 1.00e-06 ***
weekday_is_friday1             -0.20542    0.05945  -3.455 0.000558 ***
global_rate_negative_words     -4.03722    1.58745  -2.543 0.011034 *
min_positive_polarity          -0.56888    0.25074  -2.269 0.023353 *
avg_negative_polarity          -0.27273    0.13208  -2.065 0.039022 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8438 on 2978 degrees of freedom
Multiple R-squared:  0.1512,    Adjusted R-squared:  0.1452
F-statistic: 25.26 on 21 and 2978 DF,  p-value: < 2.2e-16
```

Graph 9: Correlation plot for the logarithms of all variables

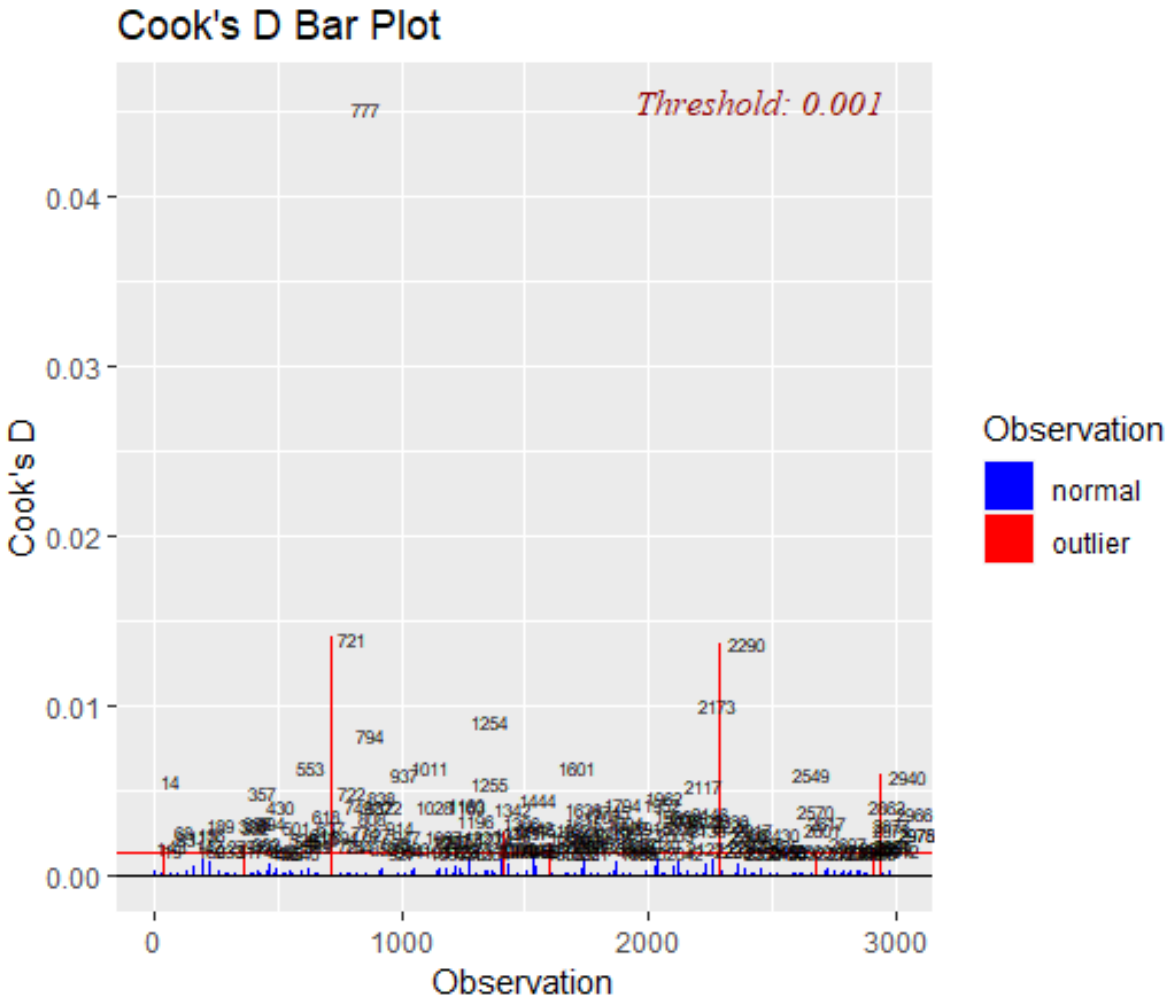


Graph 10: rstudent plot, for equality of variance test (homoscedasticity)



Graph 11: Outliers using 3x mean cooks distance.

Threshold: 3 * mean(Vector of cook's distance values)



9. References

[1] Caiyun, Wenjie, Yuqing, Ying, Fannv, Chensi. CIT(2017). Predicting the Popularity of Online News Based on Multivariate Analysis

[2] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. 17th Portuguese Conference on Artificial Intelligence(2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

[3] Deshpande, Dhanashree. ICCUBEA(2017).Prediction & evaluation of online news popularity using machine intelligence.

[4] Choudhary, Swati, Angkirat Singh Sandhu, and Tribikram Pradhan. Singapore(2017).Genetic algorithm based correlation enhanced prediction of online news popularity.

[5] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias, Serge Fdida. SNAM(2014). From Popularity Prediction to Ranking Online News

[6] K. Fernandes, P. Vinagre and P. Cortez. EPIA 2015 . A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News.