Online News Popularity Analysis

Online News from Mashable.com in the Past Two Years

Agenda

A **Introduction**

A **Descriptive Analysis**

B **Associations between Variables**

B **Predictive Analysis**

B **Conclusions**

Part A. Introduction

Objectives

- Explore features of the data
- Summarize findings from Visualizations

Data Collection

- www.kaggle.com
- mashable.com in the past two years

Sample

- Random sample
- Population of interest: content of all the articles published in the last two years from Mashable (one of the largest news websites)

Descriptions of Variables

url	URL of the article (non-predictive)
timedelta	Days between the article publication and the dataset acquisition (non-predictive)
n_tokens_title	Number of words in the title
n_tokens_content	Number of words in the content
n_unique_tokens	Rate of unique words in the content
n_non_stop_words	Rate of non-stop words in the content
n_non_stop_unique_tokens	Rate of unique non-stop words in the content
num_hrefs	Number of links
num_self_hrefs	Number of links to other articles published by Mashable
num_imgs	Number of images
num_videos	Number of videos
average_token_length	Average length of the words in the content
num_keywords	Number of keywords in the metadata

Descriptions of Variables

data_channel_is_lifestyle	Is data channel 'Lifestyle'?
data_channel_is_entertainment	Is data channel 'Entertainment'?
data_channel_is_bus	Is data channel 'Business'?
data_channel_is_socmed	Is data channel 'Social Media'?
data_channel_is_tech	Is data channel 'Tech'?
data_channel_is_world	Is data channel 'World'?
kw_min_min	Worst keyword (min. shares)
kw_max_min	Worst keyword (max. shares)
kw_avg_min	Worst keyword (avg. shares)
kw_min_max	Best keyword (min. shares)
kw_max_max	Best keyword (max. shares)
kw_avg_max	Best keyword (avg. shares)
kw_min_avg	Avg. keyword (min. shares)
kw_max_avg	Avg. keyword (max. shares)
kw_avg_avg	Avg. keyword (avg. shares)

Descriptions of Variables

self_reference_min_shares	Min. shares of referenced articles in Mashable
self_reference_max_shares	Max. shares of referenced articles in Mashable
self_reference_avg_sharess	Avg. shares of referenced articles in Mashable
weekday_is_monday	Was the article published on a Monday?
weekday_is_tuesday	Was the article published on a Tuesday?
weekday_is_wednesday	Was the article published on a Wednesday?
weekday_is_thursday	Was the article published on a Thursday?
weekday_is_friday	Was the article published on a Friday?
weekday_is_saturday	Was the article published on a Saturday?
weekday_is_sunday	Was the article published on a Sunday?
is_weekend	Was the article published on the weekend?
LDA_00	Closeness to LDA topic 0
LDA_01	Closeness to LDA topic 1
LDA_02	Closeness to LDA topic 2
LDA_03	Closeness to LDA topic 3
LDA_04	Closeness to LDA topic 4

Descriptions of Variables

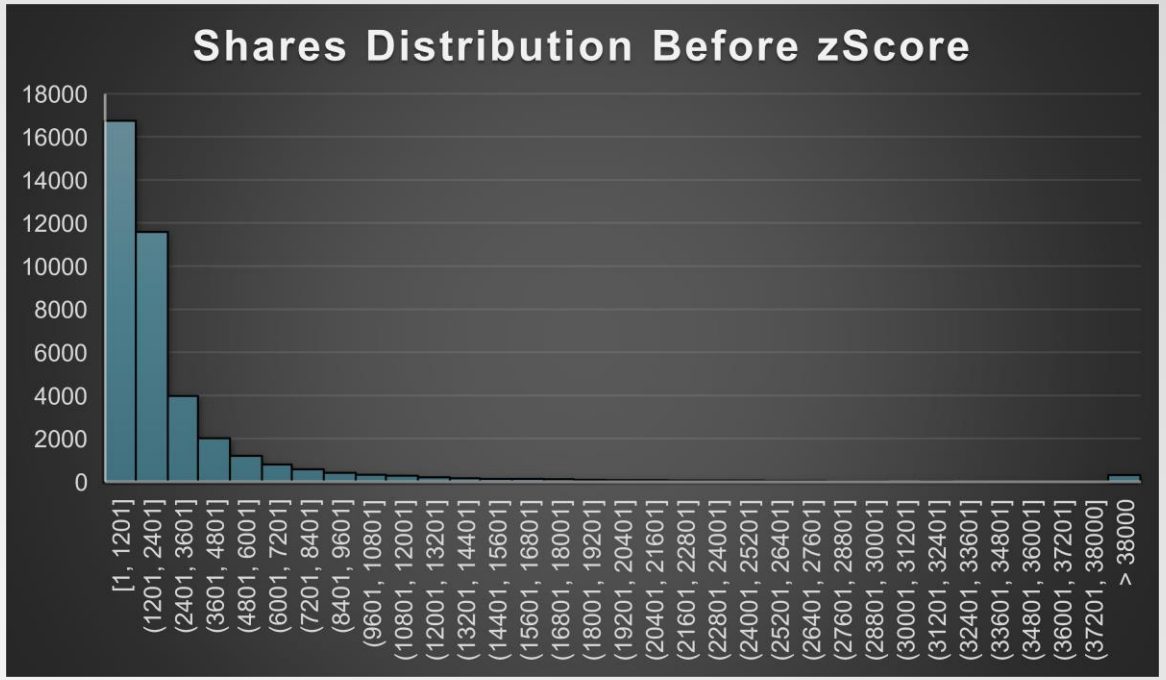
global_subjectivity	Text subjectivity
global_sentiment_polarity	Text sentiment polarity
global_rate_positive_words	Rate of positive words in the content
global_rate_negative_words	Rate of negative words in the content
rate_positive_words	Rate of positive words among non-neutral tokens
rate_negative_words	Rate of negative words among non-neutral tokens
avg_positive_polarity	Avg. polarity of positive words
min_positive_polarity	Min. polarity of positive words
max_positive_polarity	Max. polarity of positive words
avg_negative_polarity	Avg. polarity of negative words
min_negative_polarity	Min. polarity of negative words
max_negative_polarity	Max. polarity of negative words
title_subjectivity	Title subjectivity
title_sentiment_polarity	Title polarity
abs_title_subjectivity	Absolute subjectivity level
abs_title_sentiment_polarity	Absolute polarity level
shares	Number of shares (target)

Descriptions of Variables Analyzed

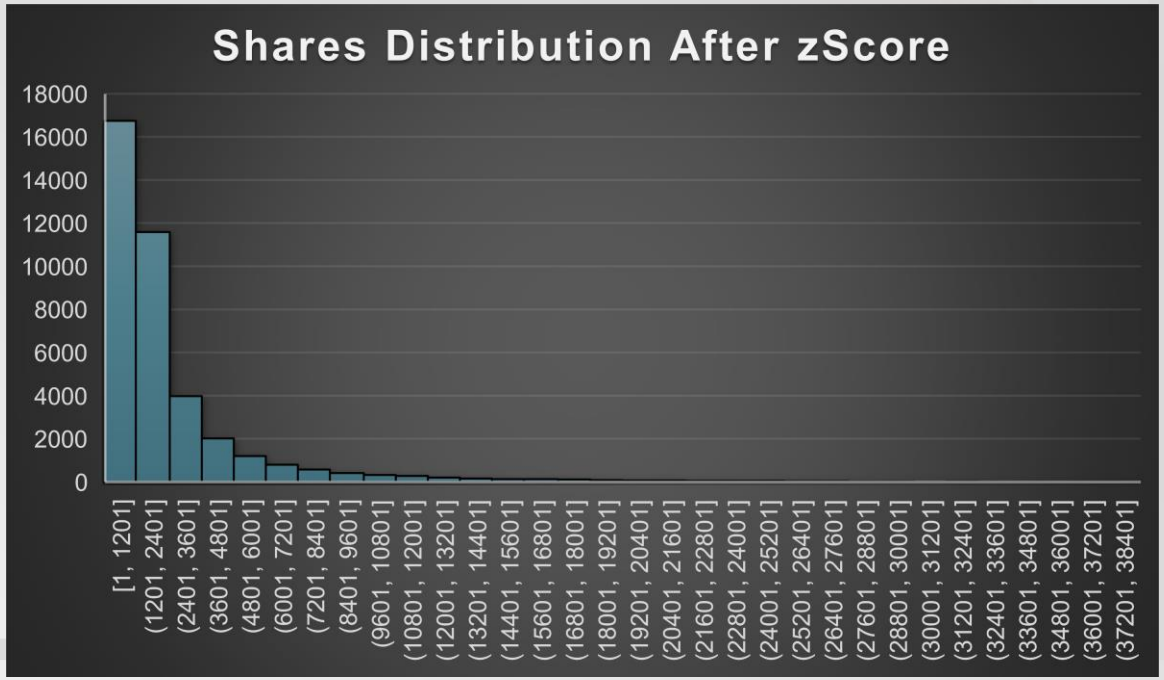
Attributes	Data type	Description
num_hrefs	Quantitative-Discrete	Number of links
num_self_hrefs	Quantitative-Discrete	Number of links to other articles published by Mashable
num_imgs	Quantitative-Discrete	Number of images
num_videos	Quantitative-Discrete	Number of videos
average_token_length	Quantitative-Continuous	Average length of the words in the content
num_keywords	Quantitative-Discrete	Number of keywords in the metadata
channel category	Qualitative- Nominal	The type of data channel
self_reference_avg_sharess	Quantitative-Continuous	Avg. shares of referenced articles in Mashable
weekday	Qualitative- Nominal	The weekday the article published
global_rate_positive_words	Quantitative-Continuous	Rate of positive words in the content
global_rate_negative_words	Quantitative-Continuous	Rate of negative words in the content
shares	Quantitative-Discrete	Number of shares (target)

Part A. Descriptive Analysis

Removing Outliers – zScore

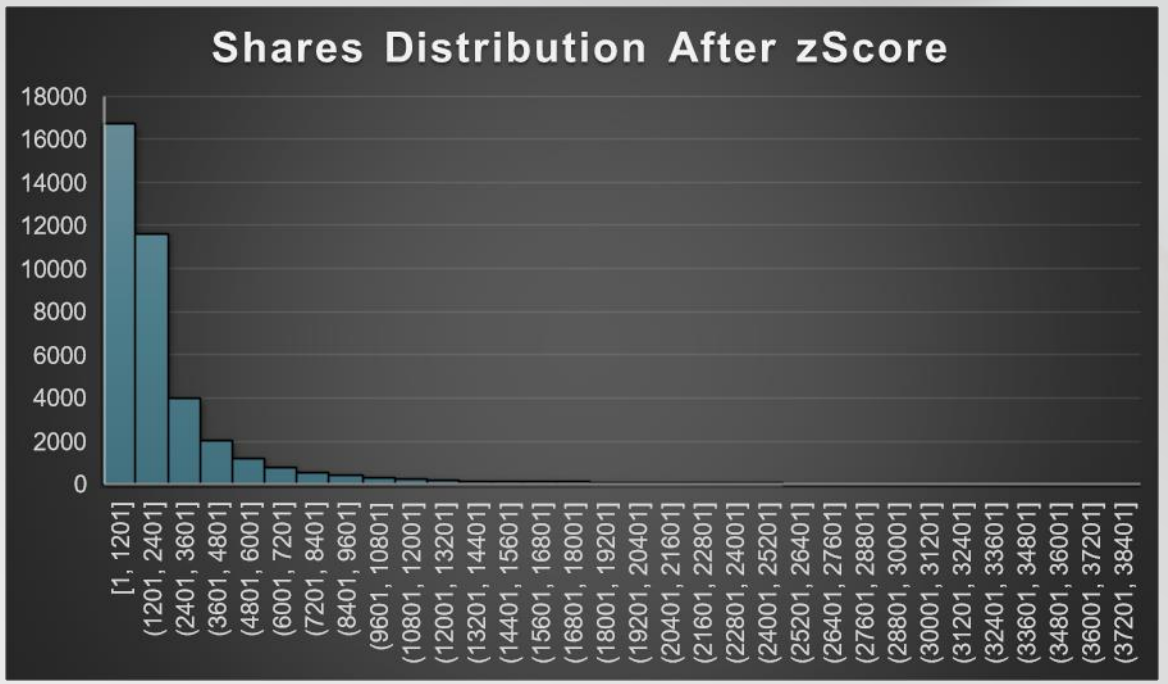


Removed 308 outliers (zScore > 3)

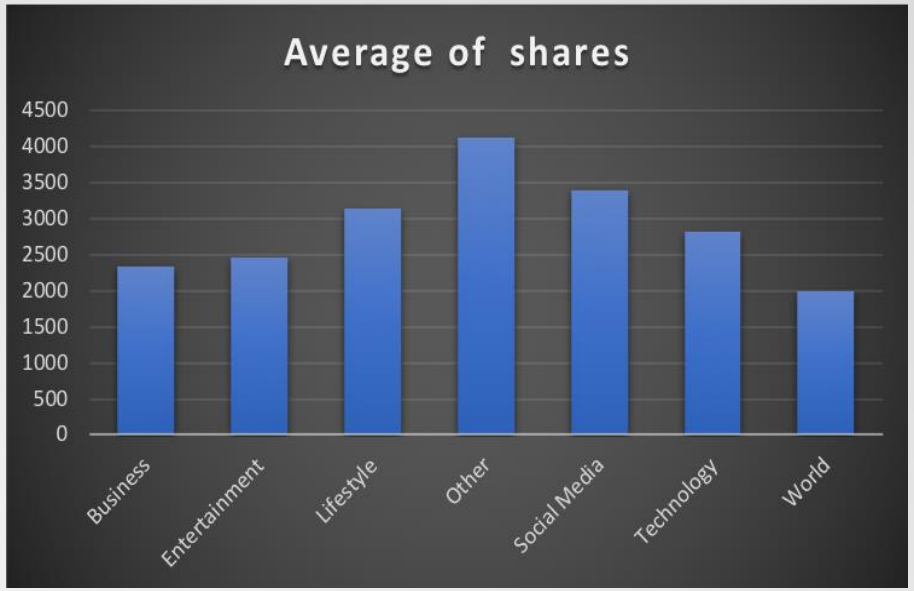


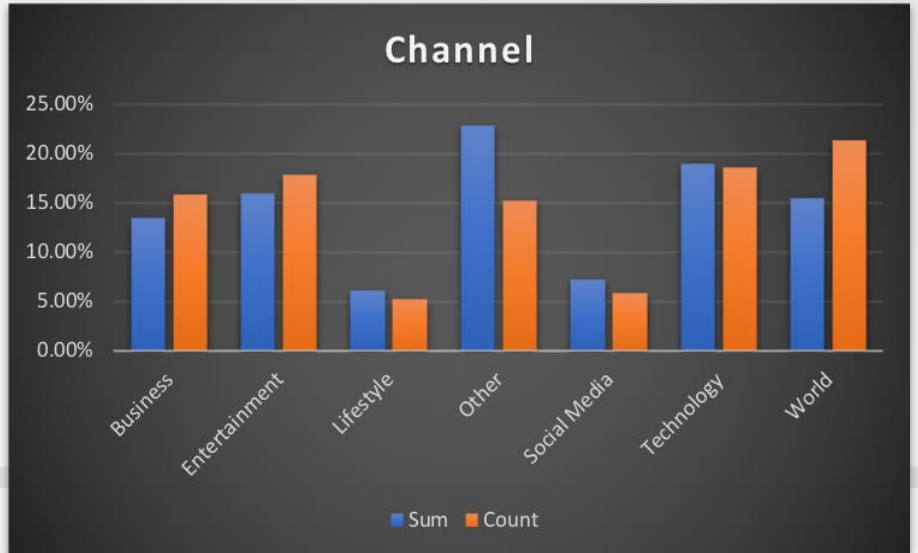
Number of Shares Analysis

Mean	2755.41626
Standard Error	19.9142389
Median	1400
Mode	1100
Standard Deviation	3949.65179
Sample Variance	15599749.3
Kurtosis	21.7064814
Skewness	4.14960847
Range	38199
Minimum	1
Maximum	38200
Sum	108387054
Count	39336



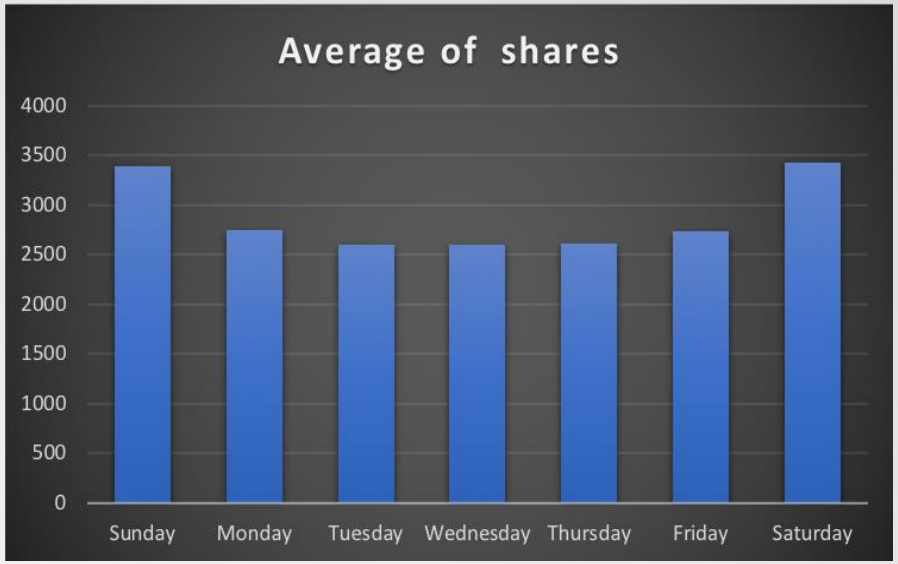
Analysis of Shares by Channels

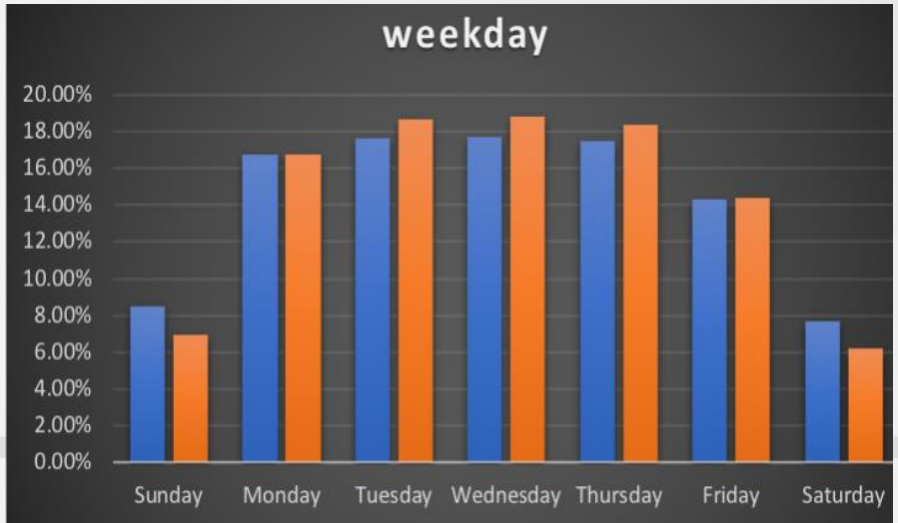




Row Labels	Count of shares	Average of shares	Sum of shares
Business	6220	2340.734727	14559370
Entertainment	7006	2468.331145	17293128
Lifestyle	2084	3149.701056	6563977
Other	6000	4135.049	24810294
Social Media	2313	3388.913532	7838557
Technology	7322	2810.870391	20581193
World	8391	1995.058396	16740535
(blank)			
Grand Total	**39336**	**2755.41626**	**108387054**

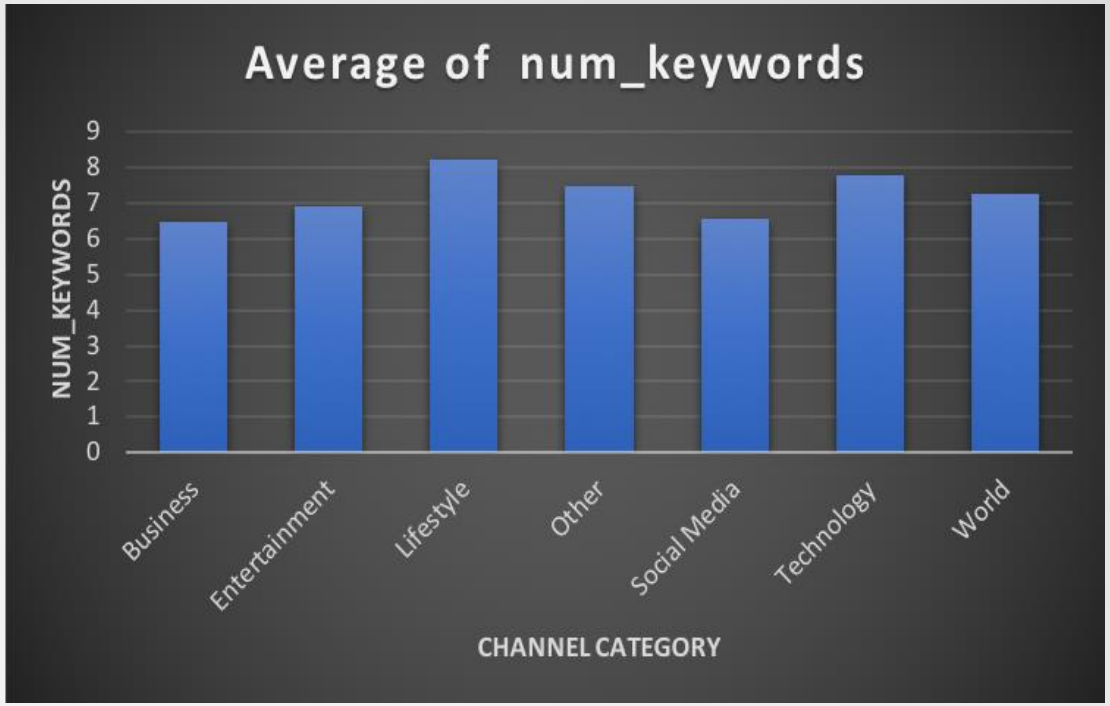
Analysis of Shares by Days of Week





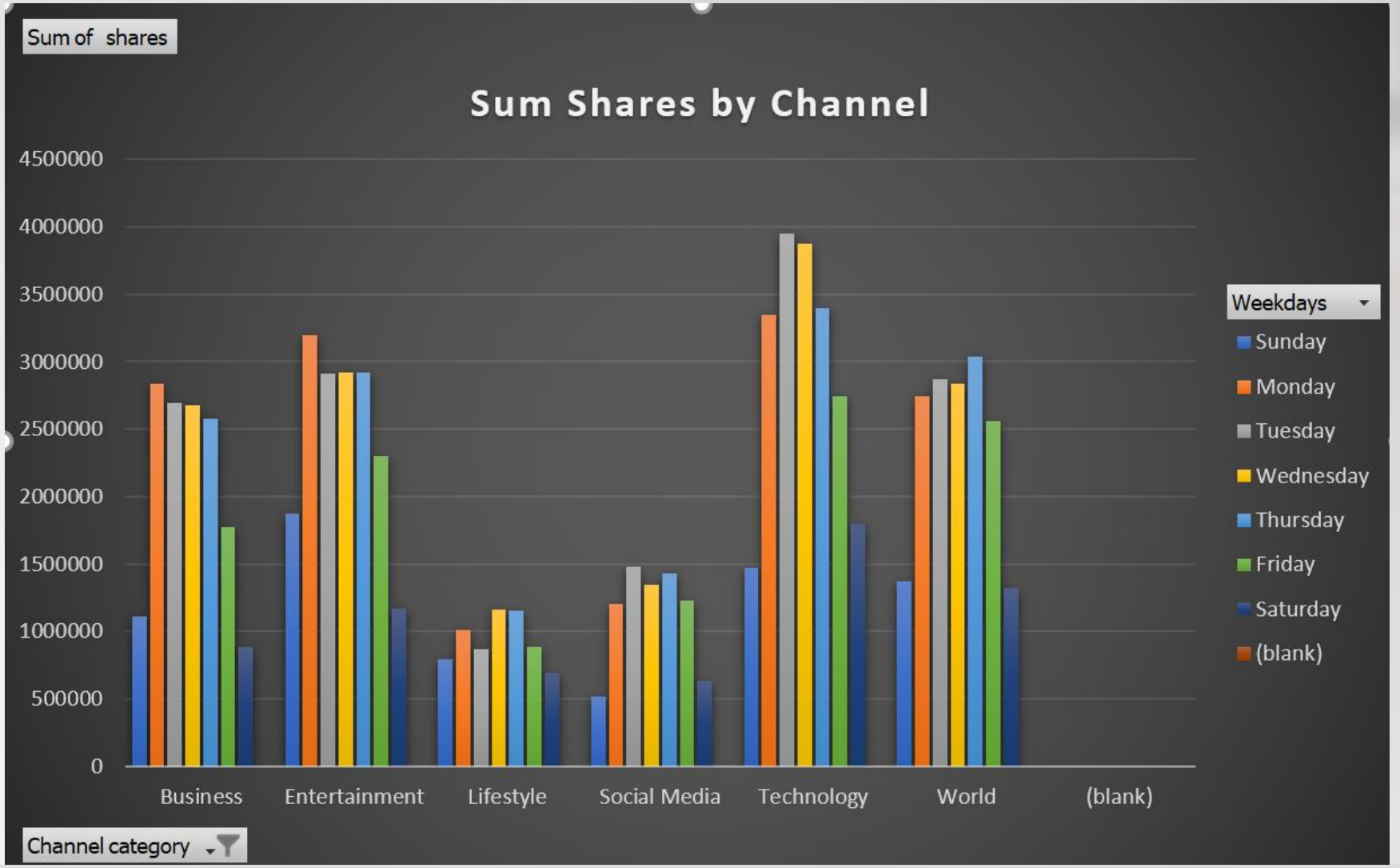
Row Labels	Sum of shares	Count of shares	Average of shares
Sunday	9209329	2719	3387.027951
Monday	18163439	6595	2754.122669
Tuesday	19083581	7332	2602.779733
Wednesday	19184819	7382	2598.864671
Thursday	18899780	7215	2619.512128
Friday	15506718	5657	2741.155736
Saturday	8339388	2436	3423.394089
(blank)			
Grand Total	**108387054**	**39336**	**2755.41626**

Number of Keywords by Channels



Row Labels	Sum of num_keywords	Average of num_keywords	Count of num_keywords
Business	40345	6.486334405	6220
Entertainment	48509	6.923922352	7006
Lifestyle	17149	8.228886756	2084
Other	44847	7.4745	6000
Social Media	15154	6.551664505	2313
Technology	56930	7.775198033	7322
World	61146	7.287093314	8391
(blank)			
Grand Total	**284080**	**7.221883262**	**39336**

Shares by Days of Week and Channels



Part A Summary

- Channels (Excluding “Others”)
  - Most of the news were published in Channel “World”, but on average, news in “World” were shared the least times
  - News is “Technology” were shared the most times
- Days of Week
  - Least amount of news were published on Weekends (Saturdays and Sundays)
  - However, on average, news published on Weekends were shared the most times.
- Number of Keywords
  - News in Channels “Lifestyle” and “Technology” had the most number of keywords.
- Channels and Days of Week
  - In the past two years, by counting the total number of shares, we find that news in “Business” and “Entertainment” were shared the most times on Monday, “Social Media” on Tuesday, “Lifestyle” on Wednesday, “Technology” and “World” on Thursday.

Part B. Predictive Analysis

Objectives

- Stating the dependent and independent variables
- Understanding how the explanatory variables are correlated with each other and with the dependent variable
- Performing multiple regression analysis to predict the dependent variable (shares)
- Determining the best predictors and interpreting the results

Dependent & Independent Variables

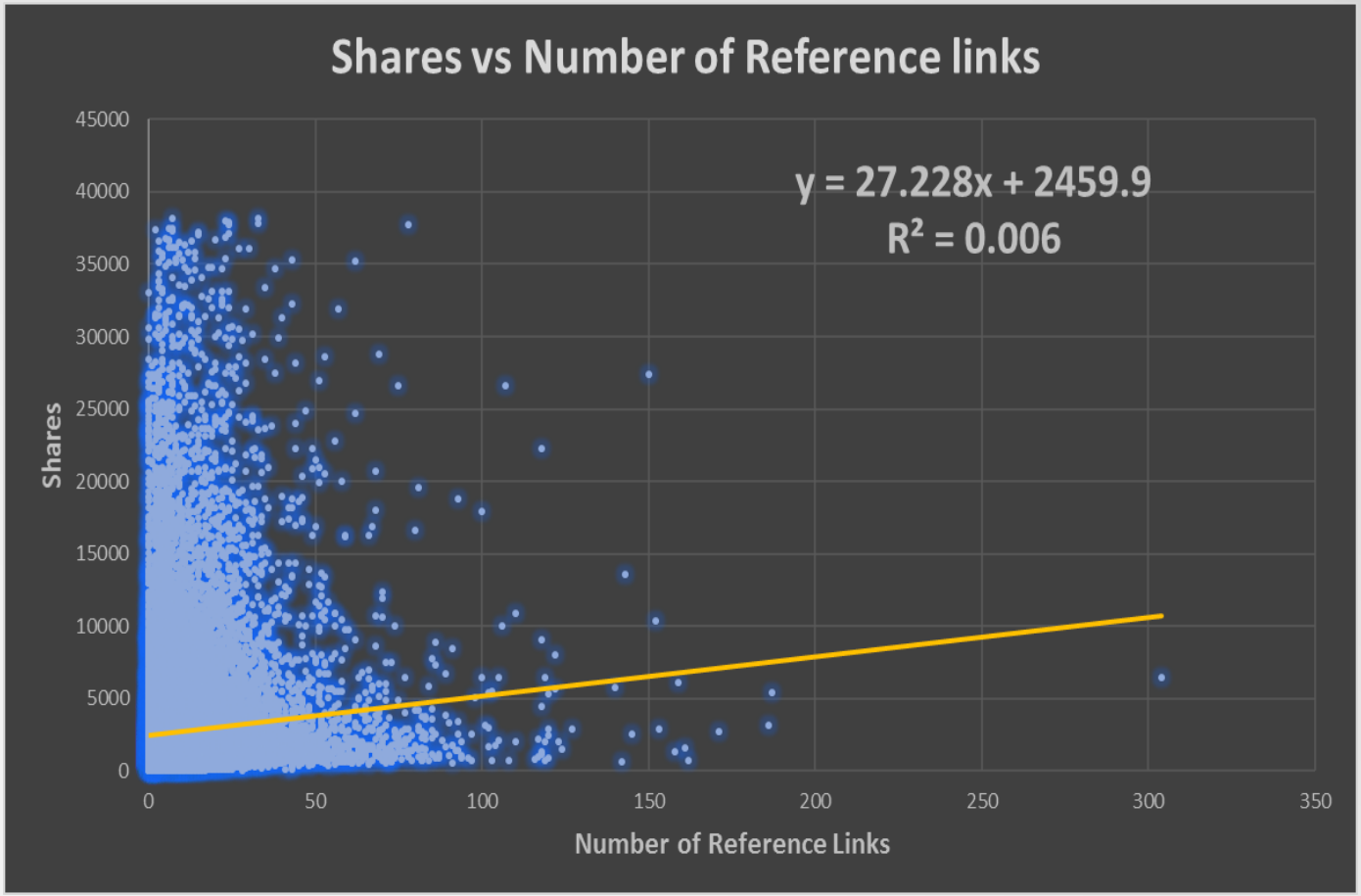
- **Dependent Variable:** Shares

- **Independent/Explanatory Variables:** num_hrefs, num_self_hrefs, num_imgs, num_videos, average_token_length, num_keywords, Channel category (Entertainment), Channel category (Business), Channel category (Social Media), Channel category (Technology), Channel category (Lifestyle), Channel category (World), Channel category (Other), self_reference_avg_shares, weekday_is_Monday, weekday_is_Tuesday, weekday_is_Wednesday, weekday_is_Thursday, weekday_is_Friday, weekday_is_Saturday, weekday_is_Sunday, global_rate_positive_words, global_rate_negative_words

Expectations

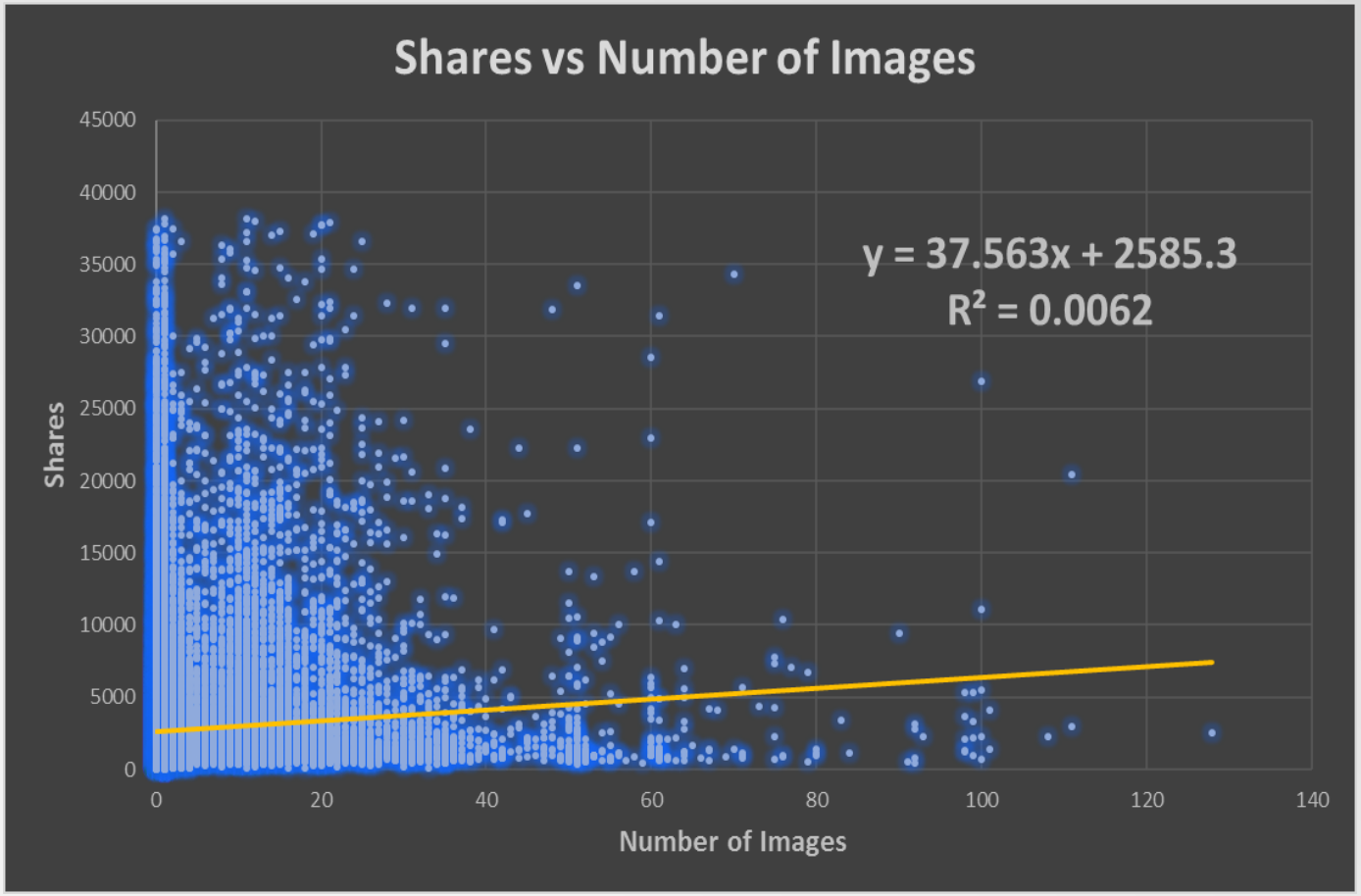
- **Num_href:** This refers to the number of links. We expect that, as the number of news links referred increases, shares would also increase. This is because we expect people would appreciate having more sources/links to obtain additional piece of information about the news articles, thereby increasing its shares
- **Num_images:** This refers to the number of images in the news article. We expect that, as this increases, shares would also increase, as they say, "*a picture is worth a thousand words*"
- **Num_videos:** This refers to the number of videos in the news article. We expect that, as this increases, shares would also increase
- **Global_rate_positive_words:** This refers to the rate of positive words used in the news article. We expect that, as this increases, shares would also increase. We expect people prefer positivity in the news.
- **Global_rate_negative_words:** This refers to the rate of negative words used in the news article. We expect that, as this increases, shares would decrease. We expect negativity could potentially drive away people from the news article.

Shares vs Number of links



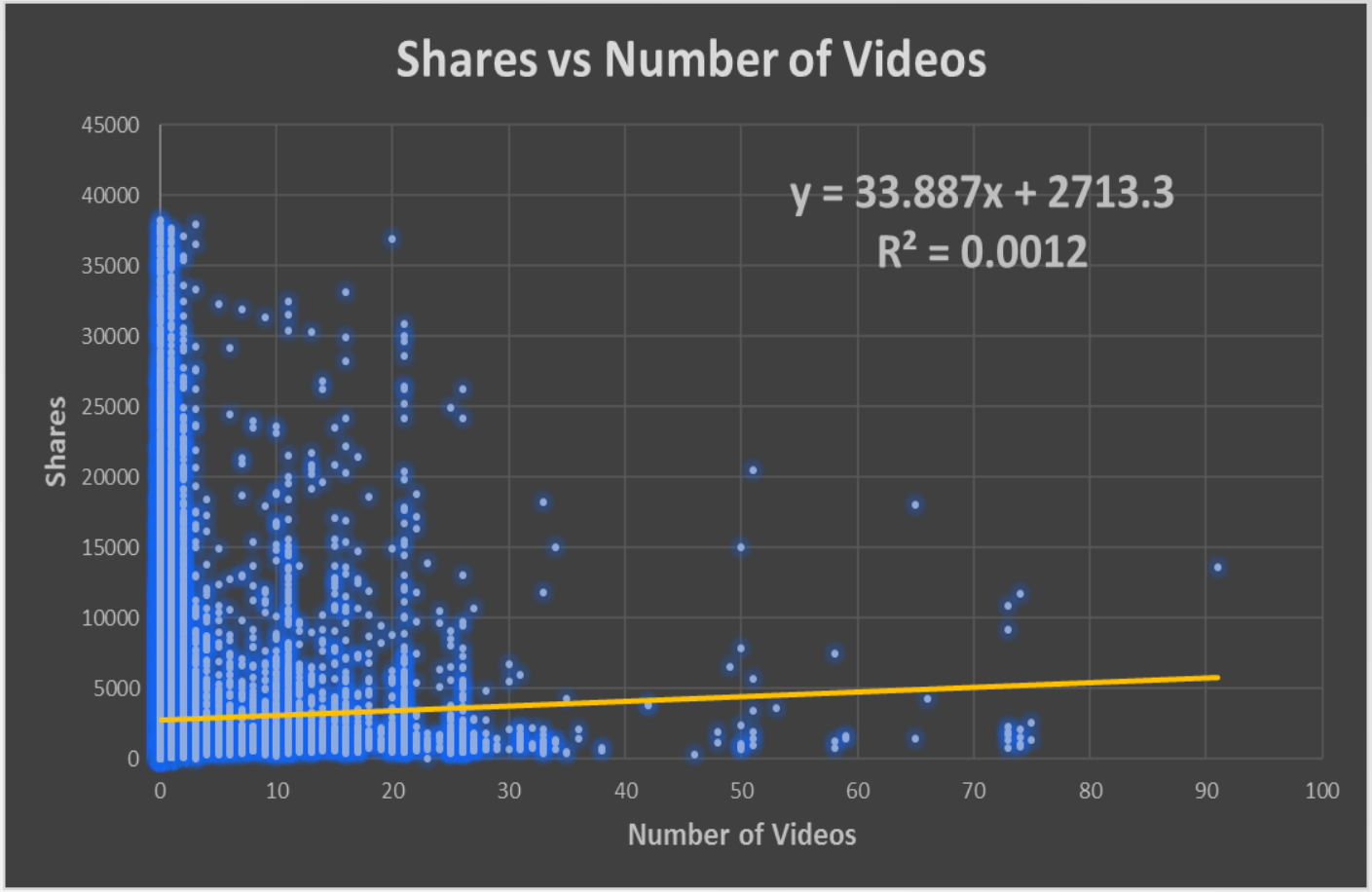
- From the scatter plot we see that, there is a positive trend between shares and number of links
- If a news article has a greater number of links, the shares for that news will increase

Shares vs Number of Images



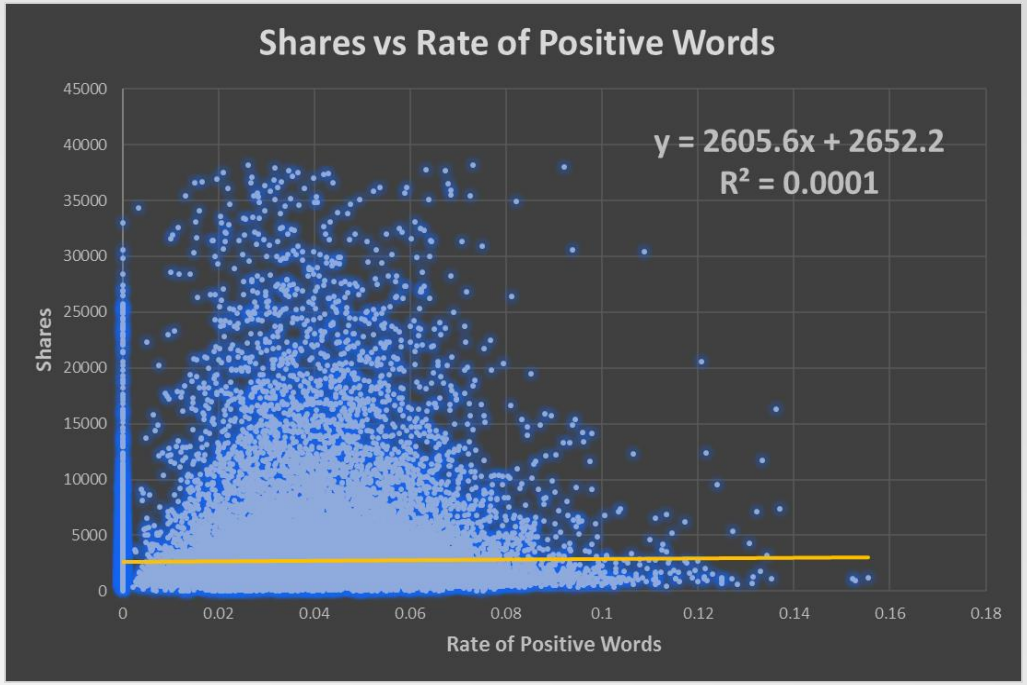
- From the scatter plot we see that, there is a positive trend between shares and number of images
- If a news article has a many images, the shares for that news will increase

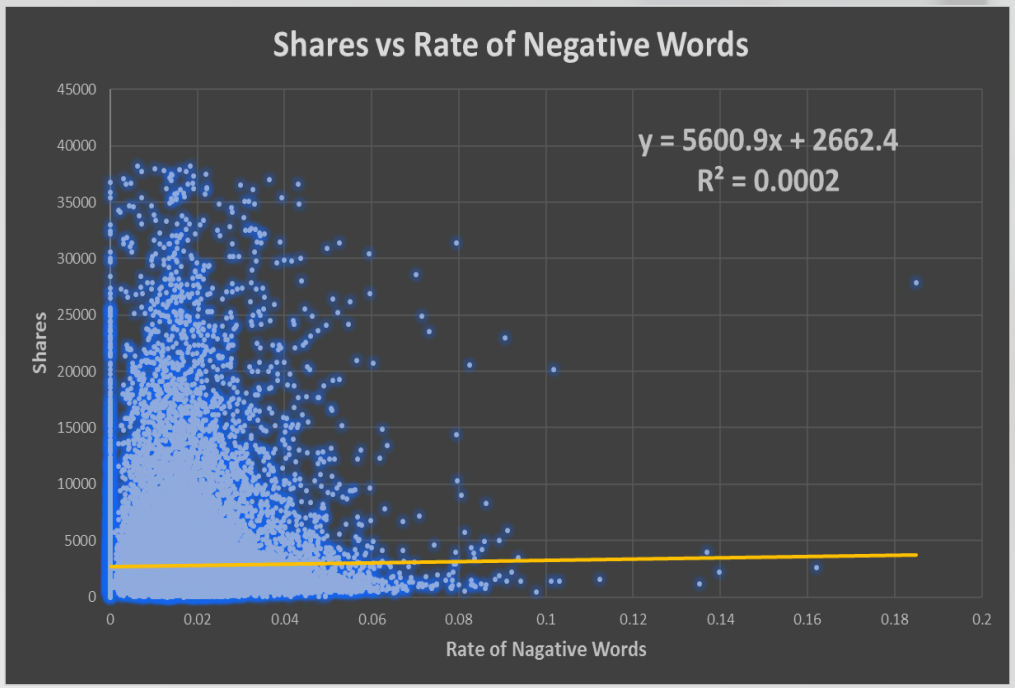
Shares vs Number of Videos



- From the scatter plot we see that, there is a positive relation between shares and number of videos
- If a news article has a many videos, the shares for that news will increase

Shares vs Rate of positive & negative words





- From the scatter plots we see that, there is almost no trend between shares and rate of positive words.

- Surprisingly, there is slight positive trend between rate of negative words and the shares.

Correlation Analysis

Linear Correlation Table	num_hrefs NewsPop	num_imgs NewsPop	num_videos NewsPop	global_rate_positive_words NewsPop	global_rate_negative_words NewsPop	shares NewsPop
num_hrefs	1.000					
num_imgs	0.343	1.000				
num_videos	0.115	-0.066	1.000			
global_rate_positive_words	0.056	-0.041	0.072	1.000		
global_rate_negative_words	0.032	0.025	0.179	0.107	1.000	
shares	0.078	0.079	0.035	0.011	0.015	1.000

- Number of links is positively correlated with the Shares. As the number of links increases, the shares will also increase. This matches with our expectations.
- Number of images is positively correlated with the Shares. As the number of images on the news article increases, the shares will also increase. This matches with our expectations.
- Number of videos is positively correlated with the Shares. As the number of videos on the news article increases, the shares will also increase. This matches with our expectations.

Correlation Analysis

Linear Correlation Table	num_hrefs NewsPop	num_imgs NewsPop	num_videos NewsPop	global_rate_positive_words NewsPop	global_rate_negative_words NewsPop	shares NewsPop
num_hrefs	1.000					
num_imgs	0.343	1.000				
num_videos	0.115	-0.066	1.000			
global_rate_positive_words	0.056	-0.041	0.072	1.000		
global_rate_negative_words	0.032	0.025	0.179	0.107	1.000	
shares	0.078	0.079	0.035	0.011	0.015	1.000

Global rate of positive words is positively correlated with Shares, although the correlation is very low. The low correlation score is probably the reason why it was difficult to see a positive trend in our scatter plot. As the rate of positive words increases, the shares will also increase. Although the correlation is low, it still matches with our expectations.

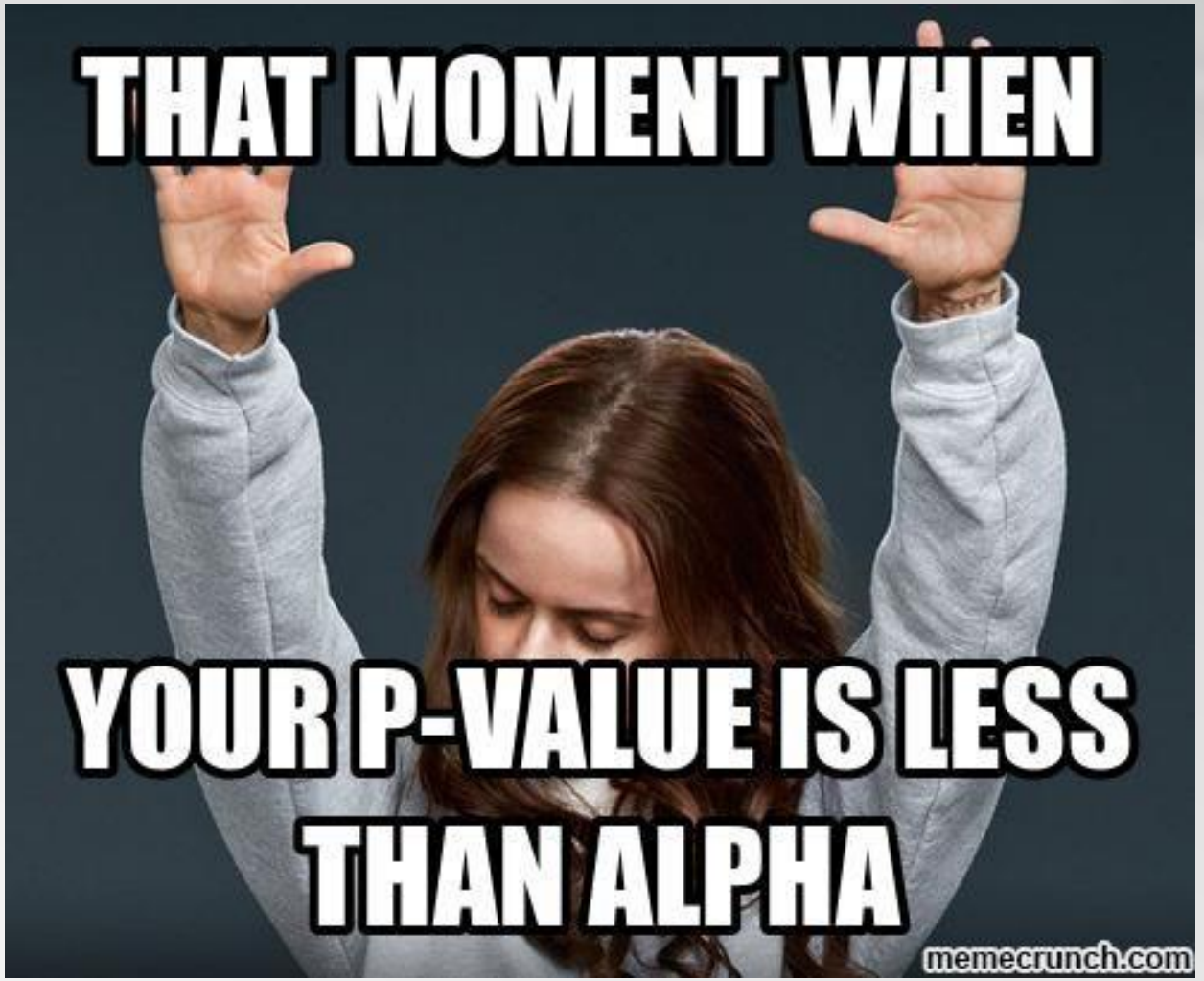
Global rate of negative words is positively correlated with Shares, although the correlation is very low. As the rate of negative words increases, the shares will also increase. This does not match with our expectations.

Correlation Analysis

Correlation Analysis for all the explanatory variables:

	num_hrefs	num_self_hrefs	num_imgs	num_videos	average_token_length	num_keywords	Entertainment	Business	Social Media	Technology	Lifestyle	World	Other	self_reference_avg_sharess	weekday_is_monday	weekday_is_tuesday	weekday_is_wednesday	weekday_is_thursday	weekday_is_friday	weekday_is_saturday	weekday_is_sunday	global_rate_positive_words	global_rate_negative_words	shares
Linear Correlation Table	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop	NewsPop
num_hrefs	1.000																							
num_self_hrefs	0.396	1.000																						
num_imgs	0.343	0.240	1.000																					
num_videos	0.115	0.078	-0.066	1.000																				
average_token_length	0.223	0.126	0.033	-0.003	1.000																			
num_keywords	0.126	0.100	0.088	-0.022	-0.017	1.000																		
Channel category (Entertai	-0.008	0.025	0.099	0.147	-0.040	-0.073	1.000																	
Channel category (Busines	-0.058	-0.055	-0.142	-0.065	0.072	-0.167	-0.202	1.000																
Channel category (Social M	0.052	0.093	-0.007	-0.008	0.025	-0.088	-0.116	-0.108	1.000															
Channel category (Technol	-0.063	0.166	-0.005	-0.093	0.019	0.139	-0.223	-0.207	-0.120	1.000														
Channel category (Lifestyle	0.053	-0.048	0.011	-0.045	0.011	0.125	-0.110	-0.103	-0.059	-0.113	1.000													
Channel category (World)	-0.031	-0.119	-0.107	-0.088	0.081	0.018	-0.242	-0.226	-0.130	-0.249	-0.123	1.000												
Channel category (Other)	0.104	-0.045	0.164	0.144	-0.165	0.056	-0.197	-0.184	-0.106	-0.203	-0.100	-0.221	1.000											
self_reference_avg_shares	0.025	0.023	0.021	0.035	0.040	0.003	-0.026	-0.006	0.025	0.018	-0.003	-0.048	0.055	1.000										
weekday_is_monday	-0.011	0.002	-0.004	0.006	0.002	-0.008	0.001	-0.001	-0.006	0.009	-0.009	0.000	-0.001	-0.002	1.000									
weekday_is_tuesday	-0.011	-0.026	-0.012	-0.002	0.010	-0.009	-0.009	-0.003	-0.007	-0.006	-0.001	0.018	0.004	0.006	-0.215	1.000								
weekday_is_wednesday	-0.004	-0.017	0.007	-0.011	0.003	0.007	0.004	0.000	0.000	-0.016	0.018	0.008	-0.007	-0.006	-0.216	-0.230	1.000							
weekday_is_thursday	0.011	0.003	0.002	0.008	0.000	0.002	0.013	0.002	0.001	0.003	-0.001	-0.016	-0.002	-0.005	-0.213	-0.227	-0.228	1.000						
weekday_is_friday	0.008	0.029	0.002	-0.001	-0.015	0.003	-0.014	-0.004	0.009	0.013	-0.001	-0.009	0.010	-0.007	-0.184	-0.196	-0.197	-0.194	1.000					
weekday_is_saturday	0.004	0.016	0.001	0.002	-0.007	0.011	0.011	-0.005	0.000	0.001	-0.001	-0.008	0.002	0.009	-0.115	-0.123	-0.123	-0.122	-0.105	1.000				
weekday_is_sunday	0.008	0.003	0.008	-0.001	0.004	-0.002	-0.006	0.013	0.004	-0.002	-0.010	0.003	-0.006	0.012	-0.122	-0.130	-0.131	-0.129	-0.112	-0.070	1.000			
global_rate_positive_words	0.056	0.121	-0.041	0.072	0.322	0.051	0.023	0.089	0.101	0.090	0.064	-0.248	-0.035	0.010	-0.005	0.003	0.004	-0.006	0.006	0.007	-0.008	1.000		
global_rate_negative_words	0.032	0.011	0.025	0.179	0.228	-0.038	0.106	-0.076	-0.020	-0.096	-0.006	0.019	0.063	0.021	-0.012	0.003	0.009	0.003	0.000	0.004	-0.008	0.107	1.000	
shares	0.078	0.003	0.079	0.035	-0.040	0.039	-0.034	-0.046	0.040	0.007	0.024	-0.100	0.148	0.067	-0.018	-0.012	0.009	0.007	0.010	0.009	-0.003	0.011	0.015	1.000

Multiple Linear Regression Model



Multiple Regression for shares *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.2050	0.0420	0.0415	3866.839532	0	0

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	21	25775599124	1227409482	82.0875274	< 0.0001
Unexplained	39314	5.87841E+11	14952447.97		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	2082.492907	159.9554571	13.01920512	< 0.0001	1768.976319	2396.009494
num_hrefs	23.09345157	2.063747282	11.1900579	< 0.0001	19.04845669	27.13844645
num_self_hrefs	-44.38267988	5.781316061	-7.676916365	< 0.0001	-55.7142	-33.05115975
num_imgs	19.44792921	2.632067305	7.38884191	< 0.0001	14.28901326	24.60684516
num_videos	12.92230761	5.059820428	2.553906368	0.0107	3.00493648	22.83967875
average_token_length	-168.2040696	26.65973268	-6.309293181	< 0.0001	-220.4577943	-115.950345
num_keywords	41.36504527	10.72152508	3.85813072	0.0001	20.35059528	62.37949525
Channel category (Business)	419.2145087	67.33095466	6.226178001	< 0.0001	287.2441995	551.1848178
Channel category (Entertainment)	365.5908573	66.12251251	5.528992221	< 0.0001	235.9891242	495.1925905
Channel category (Lifestyle)	958.3596729	96.95495472	9.884586875	< 0.0001	768.3256029	1148.393743
Channel category (Other)	1810.706741	70.15718517	25.80928435	< 0.0001	1673.196952	1948.216531
Channel category (Social Media)	1360.070989	94.3747586	14.41138509	< 0.0001	1175.094166	1545.047811
Channel category (Technology)	841.0885412	65.43211168	12.85436951	< 0.0001	712.8400105	969.3370719
self_reference_avg_sharess	0.008960971	0.000812627	11.02716824	< 0.0001	0.007368203	0.010553739
weekday_is_monday	-96.12304172	88.22852576	-1.089478045	0.2759	-269.0530986	76.80701519
weekday_is_tuesday	-47.42294023	86.94933374	-0.545408897	0.5855	-217.8457497	122.9998692
weekday_is_wednesday	123.43206	86.87767348	1.420756968	0.1554	-46.8502936	293.7144136
weekday_is_thursday	93.07042742	87.10896631	1.068436825	0.2853	-77.66526574	263.8061206
weekday_is_friday	120.7938349	90.3591701	1.336818773	0.1813	-56.31233675	297.9000066
weekday_is_saturday	168.9204672	108.0238306	1.563733356	0.1179	-42.80886879	380.6498032
global_rate_positive_words	974.7991366	1254.146527	0.777260962	0.4370	-1483.358568	3432.956841
global_rate_negative_words	5870.563901	1914.69595	3.066055423	0.0022	2117.713258	9623.414543

Multiple Linear Regression Model with Interaction Variables

- Channel Category with Rate of Negative Words

Channel category (Business) * global_rate_negative_words	14169.29545	7389.834884	1.917403524	0.0552	-314.9607731	28653.55167
Channel category (Entertainment) * global_rate_negative_words	1118.13014	5942.03374	0.188172971	0.8507	-10528.4006	12764.66088
Channel category (Lifestyle) * global_rate_negative_words	29225.03403	10573.40041	2.764014686	0.0057	8500.911889	49949.15617
Channel category (Other) * global_rate_negative_words	13723.31092	5739.035831	2.391222381	0.0168	2474.661025	24971.96083
Channel category (Social Media) * global_rate_negative_words	14392.05207	9203.465502	1.563764439	0.1179	-3646.964301	32431.06844
Channel category (Technology) * global_rate_negative_words	19414.07211	7072.185298	2.745130578	0.0061	5552.41681	33275.72741

- Coefficient
  - Lifestyle
  - Entertainment
- P-Value
  - Lifestyle
  - Technology
  - Entertainment
  - Social Media

Variables Coefficient explanation

- Continuous variable

average_token_length	-168.2040696
num_keywords	41.36504527

Variable: num_keywords

- If an article has one more keyword in its metadata, the shares of this article is predicted to increase 41.4 on average.

Variable: average_length

- If an article's average word length increases one unit, the shares of this article is predicted to decrease 168.2 on average.

Variables Coefficient explanation

- Dummy variable

Channel category (Business)	419.2145087
Channel category (Entertainment)	365.5908573
Channel category (Lifestyle)	958.3596729
Channel category (Other)	1810.706741
Channel category (Social Media)	1360.070989
Channel category (Technology)	841.0885412

- Channel World's coefficient = 0
- When an article from channel Social Media, it is predicted to have 1360 more shares than the articles from channel Worlds on average.
- When an article from channel Other, it is predicted to have 1810.7-1360.1 =450.6 more shares than the articles from channel Worlds on average.

Variables Coefficient explanation

- Interaction variable

Channel category (Business) * global_rate_negative_words	14169.29545
Channel category (Entertainment) * global_rate_negative_words	1118.13014
Channel category (Lifestyle) * global_rate_negative_words	29225.03403
Channel category (Other) * global_rate_negative_words	13723.31092
Channel category (Social Media) * global_rate_negative_words	14392.05207
Channel category (Technology) * global_rate_negative_words	19414.07211

- When an article is from channel Life style, its negative words rate in content increase 1%, it is predicted to have 29225 more shares on average.
- When an article's negative words rate in content increase 1%, if the article is from channel Life style, it is predicted to have 29225 – 1118 =28107 more shares than if it is from channel Entertainment

R-Squared Adjusted

- With Interaction Variables

Multiple Regression for shares *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.2061	0.0425	0.0418	3866.17	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	27	2.6E+10	965554016.1	64.5975	< 0.0001	
Unexplained	39308	5.9E+11	14947241.78			

- Without Interaction Variables

Multiple Regression for shares *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.2050	0.0420	0.0415	3866.84	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	21	2.6E+10	1227409482	82.0875	< 0.0001	
Unexplained	39314	5.9E+11	14952447.97			

Confidence Interval

	shares
Conf. Intervals (One-Sample)	Data Set #1
Sample Size	39336
Sample Mean	2755.42
Sample Std Dev	3949.65
Confidence Level (Mean)	**95.0%**
Degrees of Freedom	39335
Lower Limit	**2716.38**
Upper Limit	**2794.45**
Confidence Level (Std Dev)	95.0%
Degrees of Freedom	39335
Lower Limit	3922.25
Upper Limit	3977.45

We are **95%** confident that the overall average numbers of shares of online news is between **2716** and **2794**

Hypothesis Testing #1

- Test whether “number of images” significantly affects “number of shares” at 95% confidence level

- Null Hypothesis: “Number of images” does not significantly affect “number of shares”
- Alternative Hypothesis: “Number of images” significantly affect “number of shares”

- Alpha = 1-0.95 = 0.05; P-value from regression: <0.0001
- P-Value is less than Alpha, so we can reject null hypothesis at 95% confidence level

- Conclusion: We have sample evidence that X is a goo linear predictor of Y

Hypothesis Testing #2

- Test whether “rate of negative words” significantly affects “number of shares” at 95% confidence level

- Null Hypothesis: “rate of negative words” does not significantly affect “number of shares”
- Alternative Hypothesis: “rate of negative words” significantly affect “number of shares”

- Alpha = 1-0.95 = 0.05; P-value from regression: <0.0022
- P-Value is less than Alpha, so we can reject null hypothesis at 95% confidence level

- Conclusion: We have sample evidence that X is a good linear predictor of Y