Analysis of News Articles

Introduction:-

Social media is a real world phenomenon of the saying “If the tree falls on the ground and no one is around to hear it, does it make a sound?” Mashable, an independent news platform, is a part of this social media world providing the latest news articles on digital culture, social media and technology. However, not every new article is a hit, and our group aims to find what makes an article popular by analyzing various factors related to the article.

Data Set Description:-

Source : We downloaded this data from UCI machine learning repository. The link for the source is here

This data set summarizes a heterogeneous set of features about articles published by Mashable (www.mashable.com) over a period of two years. General characteristics of this data set are:

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Integer, Real
- Number of Instances: 39797
- Number of Attributes: 61
- Missing Values: No missing values

Information about the attributes that were considered for analyses:

- Day of Week: column created from pivoted rows that contains the day of the week the article was published.
- Category: column created from pivoted rows that contains the the appropriate boolean value of the category that it belonged to.
- 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_unique_tokens: Rate of unique non-stop words in the content
- num_hrefs: Number of links
- 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
- 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
- avg_negative_polarity: Avg. 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)

The final list of variables used for analysis are referred to in the questions section

How the data set was cleaned:

- Original File contained 39797 rows x61 columns
- Pivoted all days of week Boolean columns to single categorical column [Sunday-Saturday]
- Pivoted all article category Boolean columns to single categorical column [Lifestyle, entertainment, business; social media, tech, world]
- Removed all rows that have a null value using na.omit()
- Pivoted File (33509 rows × 48 columns)
- Removed all articles with O length of 'number of words in the content'
- Removed columns that are not relevant to our testing. The columns were: ['num_self_hrefs', '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_shares', 'is_weekend', 'LDA_00', 'LDA_01', 'LDA_02', 'LDA_03', 'LDA_04', 'min_positive_polarity', 'max_positive_polarity', 'min_negative_polarity', ' max_negative_polarity, 'n_non_stop _words']
- Multiplied all columns that have a range of 0.0-1.0 by 10. The columns included: ['n_non_stop_words', 'n_on_stop_unique_tokens', 'min_positive_polarity', 'max_positive_polarity', 'avg_negative_polarity', 'min_negative_polarity', 'max_negative_polarity', 'title_subjectivity', 'title_sentiment_polarity', 'abs_title_subjectivity', 'abs _title_sentiment_polarity', 'n_unique_tokens', 'global_subjectivity', 'rate_positive_words', 'rate_negative_words', 'avg_positive_polarity', 'global_sentiment_polarity', 'global_rate_positive_words', 'global_rate_negative_words']
- Final Clean File used has 32971 rows × 24 columns

Software used:

The data cleaning was done using Tableau Prep and Python and the statistical methods were analysed and graphed using R in R Studio.

Questions to be analyzed:

How do linguistic features vary across different article categories? This question will be applied to each of the following linguistic features one at a time across article category (business, lifestyle, entertainment, technology, world social media; applies for all questions using article category)

1) How many words are there in an article?
2) How many words are there in an title of an article?
3) What is an average token length in an article?
4) How is the polarity of articles distributed within a range of [-1 to 1]?
5) How is article subjectivity varying with objective and subjective statements?
6) Rate of positive words
7) Rate of negative words

How do other features vary across article categories? This question will be applied to to the following variables one at a time split across article category:

8) Number of images
9) Number of videos
10) Number of hyperlinks

What variables affect article popularity (number of shares) on Mashable? - This question will be used to answer the following

11) What is the distribution of number of shares across article category?
12) Using a linear regression model, understand what is the effect of different variables on the number of shares for an article? Variables and details listed in analysis plan

Statistical Methods

Before diving in deep for the explanation of the Statistical Methods used, we have bifurcated our statistical analysis into two parts: Descriptive Statistics and Inferential Statistics.

DESCRIPTIVE STATISTICS:

-Descriptive statistics will be used for questions 1-11. -Visualization methods such as box plots and/or descriptive statistics tables split up by article category will be used to describe and report the relationship of each of the variables with article category

INFERENTIAL STATISTICS:

i) ANOVA:

- Used for questions 1-11 at significance level = 0.05
- Along with descriptive Analysis, the use of anova showed visualizations with interesting patters.
- Using the power of inferential statistics, the NULL hypothesis and alternative hypothesis under consideration is as: Null Hypothesis: $H_0 : \mu_{business} = \mu_{lifestyle} = \mu_{entertainment} = \mu_{technology} = \mu_{social\ media} = \mu_{world}$ Alternative Hypothesis: $H_A : The\ mean\ value\ is\ not\ equal\ for\ one\ or\ more\ categories$
- P-value will be used to report and interpret results -Assumptions made and their results:
  - Independence – observations are individual articles and assumption of independence is not violated
  - Normality – large sample size hence normality assumption holds
  - Constant variance - does not hold **why?what was done to change this? did it work?**
- Factor variable – Article category
- Response variable – each variable listed in question 1-11 that will be compared across the factor variable
- We will not be adjusting for multiple testing since we are using different features for each of the listed questions

ii) LINEAR REGRESSION:

- Used for question 12
- Assumptions made in the linear regression and their results:
  - Independence: The data points (observations) represents individual articles without any stratification and hence the independence assumption holds true here in our case.
  - Constant variance: Applying log transformation to the dependent variable helped us meet this assumption, as can be seen in the scale-location plot
  - Linearity: Applying log transformation to the dependent variable helped us meet this assumption, as can be seen in the residuals vs fitted values plot
  - Normality: From the QQ plot, the normality assumption is not met. However, since there are large number of data points, we can proceed with statistical inferences using this model.
- Response variable – number of article shares
- Independent variables – number of words in article, number of words in title, average token (word) length in article, article polarity, article subjectivity, day of publication (Monday-Sunday), article category (business, lifestyle, technology, world, entertainment, social media), number of images, number of videos, number of hyperlinks

Results.

Describe your results clearly and concisely. Use graphical displays and tables to convey descriptive information about the data and the results of the analysis.

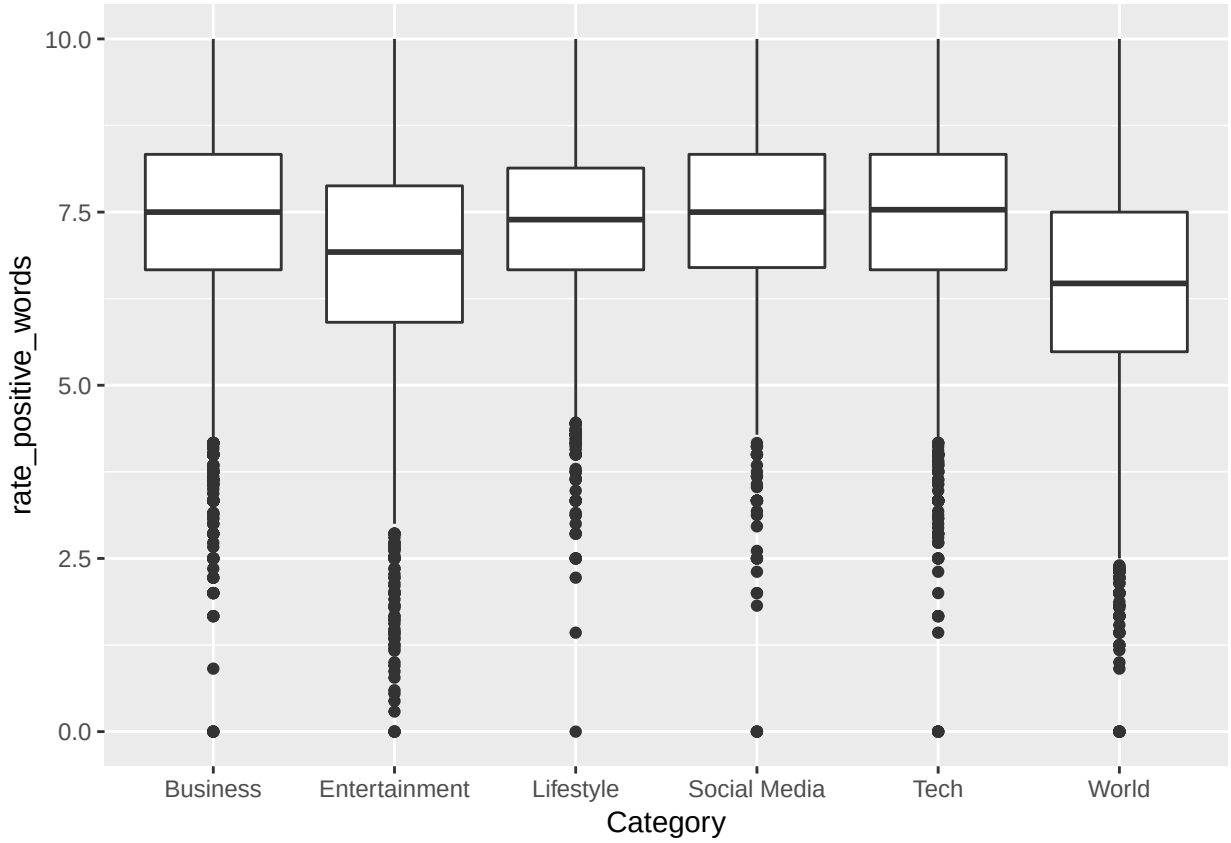
Descriptive analysis and ANOVA

add anova graphs + explanations here

- **Rate of Positive Words:**

```
data <- read.csv("Cleaned News Popularity.csv")
```

```
library("ggplot2")
ggplot(data, aes(x = Category, y = rate_positive_words)) +
  geom_boxplot()
```




```
res.aov <- aov(rate_positive_words ~ Category, data = data)

summary(res.aov)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Category        5   6105    1221   614.9 <2e-16 ***
## Residuals   32965  65464       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

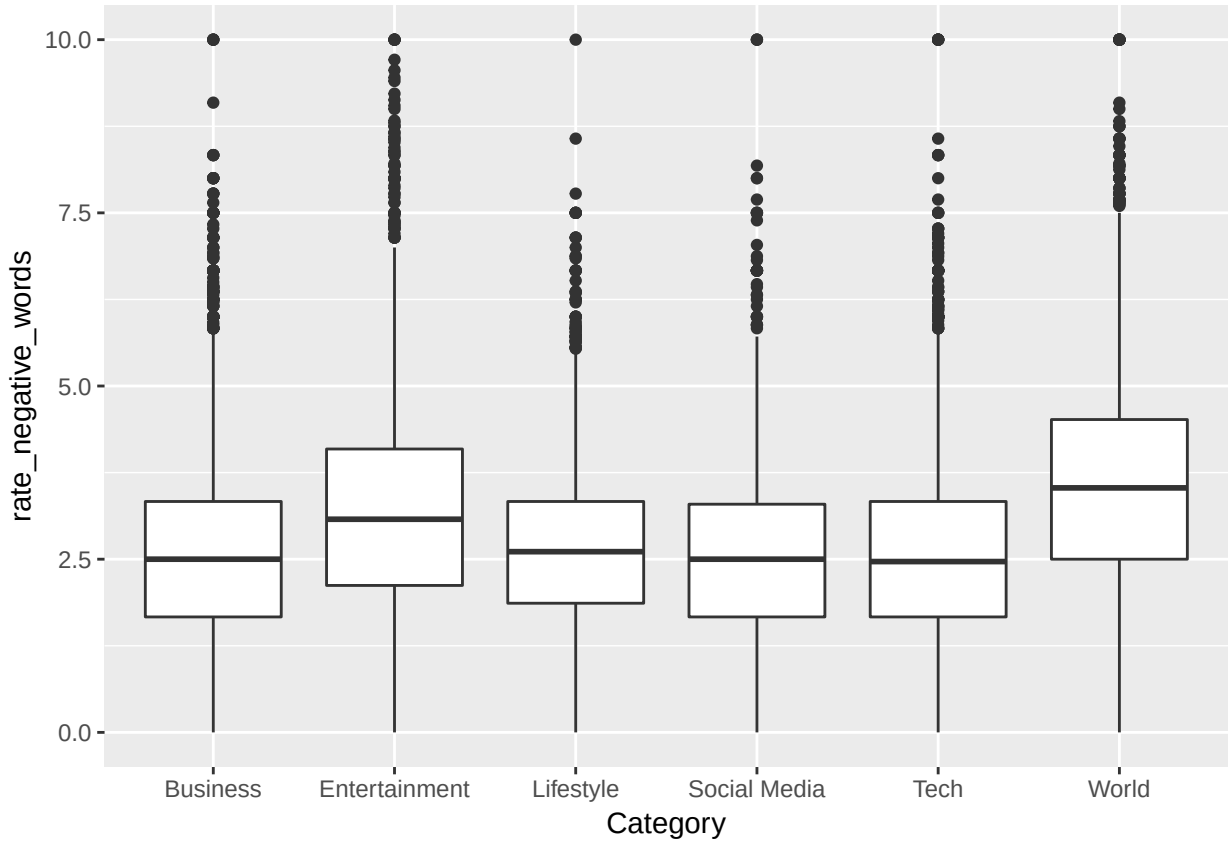
NULL Hypothesis: The mean rate of positive words across all the categories is equal.

Alternative Hypothesis: The mean rate of positive words across all the categories is NOT equal.

On performing the ANOVA test to check the significance level, we got the p value as **<2e-16**. Since the value of p is less than 0.05, we reject the NULL hypothesis.

From the box plots, we observed that the number of outliers are too high on the lower side of the inter quartile range which clearly agrees to the fact that there is a great amount of variance in the data and the variance is among the categoreis is not equal.

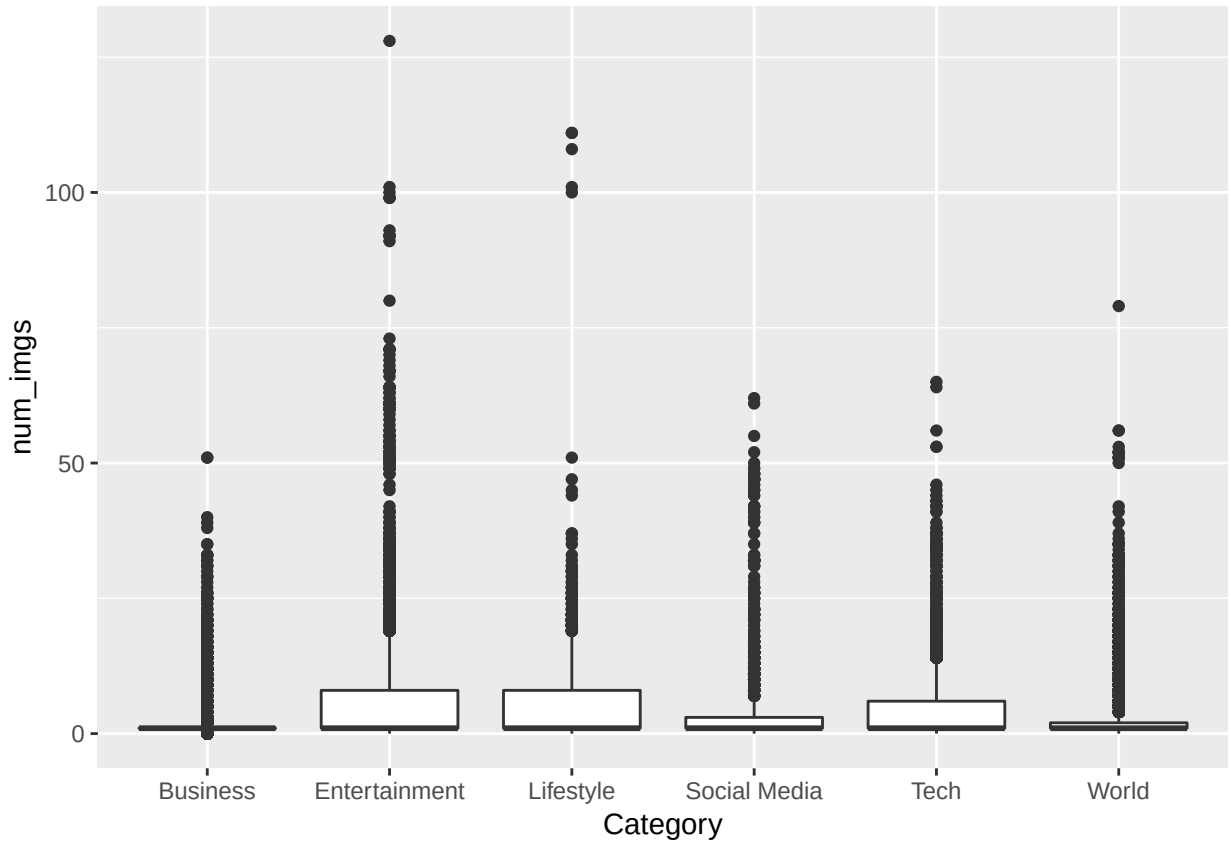
```
ggplot(data, aes(x = Category, y = rate_negative_words )) +
  geom_boxplot()
```



To add on this, when we plot the box plot for the rate of negative words, it appears to be the mirror image for the box plot for rate of positive words.This makes us to emphasize on the fact that the number of positive words and negative words follow the same distribution.

- **Number of Images**

```
ggplot(data, aes(x = Category, y=num_imgs)) +
  geom_boxplot()
```




```
res.aov <- aov(num_imgs ~ Category, data = data)

summary(res.aov)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## Category        5   85710   17142     307 <2e-16 ***
## Residuals   32965 1840393      56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NULL Hypothesis: The mean number of images across all the categories is equal.

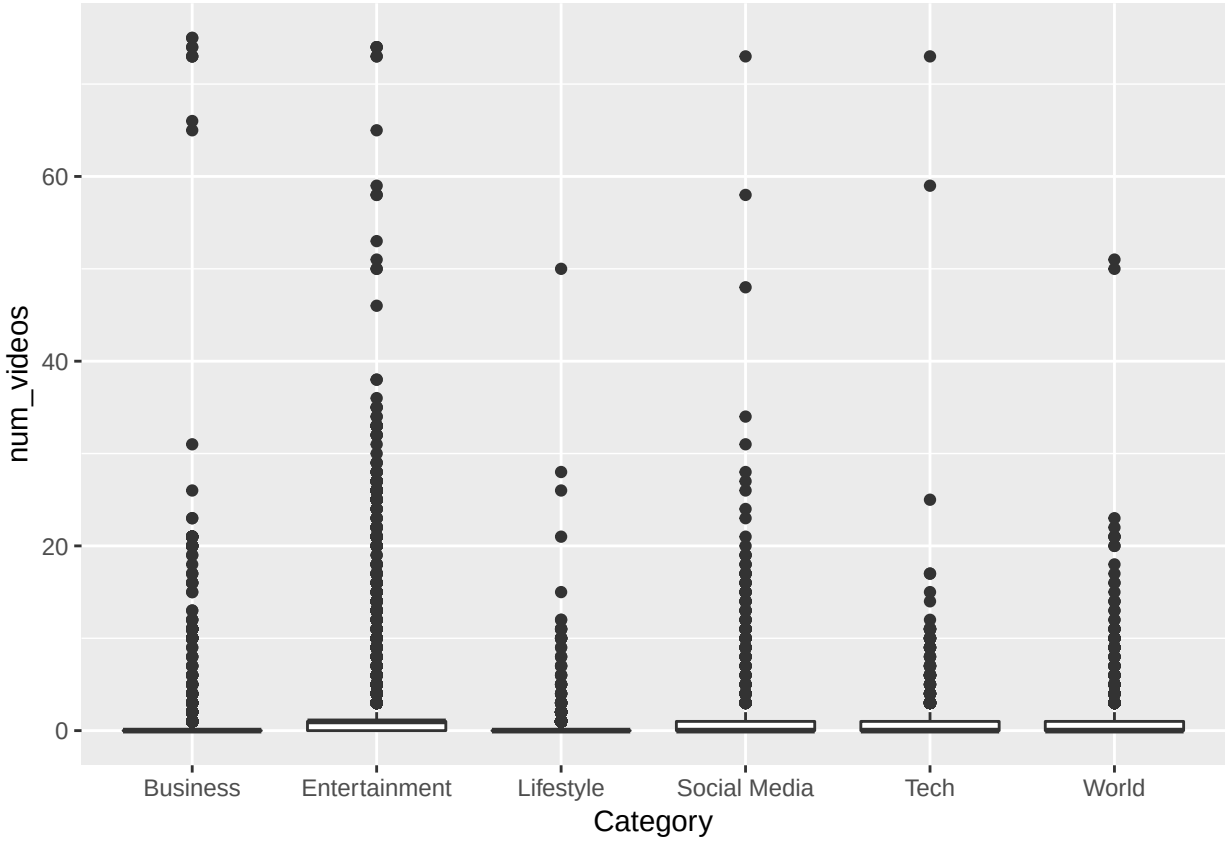
Alternative Hypothesis: The mean number of images across all the categories is NOT equal.

On performing the ANOVA test to check the significance, we got the p value as **<2e-16**. Since the value of p is less than 0.05, we reject the NULL hypothesis.

From the box plots for the category, we observe that the inter-quartile range of each category varies heavily and the distribution of the outliers above the top whisker varies significantly high for the Entertainment and Lifestyle categories.

- **Number of Videos**

```
ggplot(data, aes(x = Category, y = num_videos )) +
  geom_boxplot()
```




```
res.aov <- aov(num_videos ~ Category, data = data)

summary(res.aov)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Category        5  22658    4532   352.7 <2e-16 ***
## Residuals   32965 423571      13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NULL Hypothesis: The mean number of videos across all the categories is equal.

Alternative Hypothesis: The mean number of videos across all the categories is NOT equal.

On performing the ANOVA test to check the significance, we got the p value as **<2e-16**. Since the value of p is less than 0.05, we reject the NULL hypothesis.

From the box plots for the category, we observe that the inter-quartile range for business and lifestyle is too low. Also for Business and Entertainment categories, it is observed that there are significant number of outliers in their respective box plots. This concludes on the fact that there is a signifcant change in variance across the categories.

Linear Regression

add linear regression graphs + explanations here

Model results, parameter estimates and their significance:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.3800170	0.0446290	75.7358755	0.0000000
Day.of.WeekMonday	-0.0085025	0.0072359	-1.1750455	0.2399850

	Estimate	Std. Error	t value	Pr(>|t|)
Day.of.WeekSaturday	0.1095176	0.0098238	11.1482157	0.0000000
Day.of.WeekSunday	0.1288278	0.0095855	13.4398702	0.0000000
Day.of.WeekThursday	-0.0291252	0.0071275	-4.0863254	0.0000439
Day.of.WeekTuesday	-0.0304760	0.0071011	-4.2917548	0.0000178
Day.of.WeekWednesday	-0.0329723	0.0070828	-4.6552834	0.0000032
CategoryEntertainment	-0.0774145	0.0067839	-11.4114436	0.0000000
CategoryLifestyle	0.0385776	0.0094276	4.0920065	0.0000429
CategorySocial Media	0.1310518	0.0090197	14.5295460	0.0000000
CategoryTech	0.0549911	0.0064379	8.5418074	0.0000000
CategoryWorld	-0.0907015	0.0063923	-14.1892588	0.0000000
n_tokens_title	-0.0000129	0.0009693	-0.0133538	0.9893456
n_tokens_content	0.0000099	0.0000051	1.9286829	0.0537788
num_hrefs	0.0024822	0.0002256	11.0033895	0.0000000
num_imgs	0.0011967	0.0003052	3.9208594	0.0000884
num_videos	0.0021293	0.0005758	3.6978158	0.0002178
average_token_length	-0.0634439	0.0082883	-7.6546164	0.0000000
global_subjectivity	0.0262129	0.0026597	9.8554784	0.0000000
global_sentiment_polarity	-0.0035785	0.0025131	-1.4239686	0.1544650
title_sentiment_polarity	0.0032187	0.0008282	3.8864061	0.0001019

Table 1: Regression model coefficients

All coefficients were found to be significant other than number of tokens in title, number of tokens in content, and article polarity.

We convert the results to percentage changes and create confidence intervals for the same.

	Percentage change	Lower CI	Upper CI
(Intercept)	239792.697	196029.339	293321.200
Day.of.WeekMonday	-1.939	-5.089	1.316
Day.of.WeekSaturday	28.682	23.101	34.516
Day.of.WeekSunday	34.533	28.837	40.480
Day.of.WeekThursday	-6.486	-9.447	-3.429
Day.of.WeekTuesday	-6.777	-9.717	-3.741
Day.of.WeekWednesday	-7.311	-10.227	-4.300
CategoryEntertainment	-16.327	-18.850	-13.726
CategoryLifestyle	9.289	4.737	14.040
CategorySocial Media	35.223	29.829	40.842
CategoryTech	13.499	10.248	16.845
CategoryWorld	-18.848	-21.156	-16.473
n_tokens_title	-0.003	-0.439	0.435
n_tokens_content	0.002	0.000	0.005
num_hrefs	0.573	0.471	0.676
num_imgs	0.276	0.138	0.414
num_videos	0.491	0.231	0.753
average_token_length	-13.592	-16.764	-10.298
global_subjectivity	6.222	4.954	7.504
global_sentiment_polarity	-0.821	-1.939	0.311
title_sentiment_polarity	0.744	0.368	1.121

Table 2: Transformed model coefficients as percentage changes with confidence intervals

Discussion.

This section should briefly summarize the results and conclusions. Also describe limitations of the analyses, including limitations of the data set as well as of the statistical analyses.

Linear Regression

Notable findings from regression analysis, every conclusion must be interpreted holding all other model varaibles at a constant value:

- Impact of day of week of publication on number of shares:
  - Articles published on Friday are more popular than those published on Tuesday (6.77% less shares), Wednesday (7.3% less shares) and Thursday (6.5% less shares)
  - Articles published on weekends are more popular than those on any other weekday (34.53% more shares on Sunday and 28.62% more shares on Saturday than that of Friday on average)
  - This is interesting and is likely due to engagement being high on days when people are generally not working
- It appears that World and Entertainment are the least popular categories on Mashable and Social Media and Tech are the most popular categories on Mashable (reference category in model is Business)
  - Social Media articles are shared 35.22% more and Tech articles are shared 13.5% more than Business articles on average
- It appears articles with longer words on average are not as widely shared. An increase in average length of words in an article by 1 unit decreases the number of shares by 13.59% on average
- Articles that are subjective in nature tend to be more popular. A 1 unit increase in subjectivity score increases the number of shares by 6.22% on average
- It looks like more hyperlinks, images and videos in an article have a positive effect on number of shares. The table outlines the precentage change in shares for one unit increase in hyperlink/image/video which could be why the percentage changes are smaller
- Features related to polarity, number of words in article and title do not appear to have a meaningful impact on the number of shares

Limitations: need to write better

- Only one source of articles
- Time period of publication is 2013-2015

Due to this findings cannot be generalized to all news articles

Tables and Figures

References.

List books or articles you consulted. (It is not necessary to do a lot of background research, so the reference list should be short.) References for statistical methods used in class (e.g., t-tests, and linear regression) are not required, but references should be given for advanced methods not covered in class.

Appendices

a) Regression Model

Rationale behind log transformation of dependent variable

Below are the model plots for a linear regression model with a non-transformed dependent variable

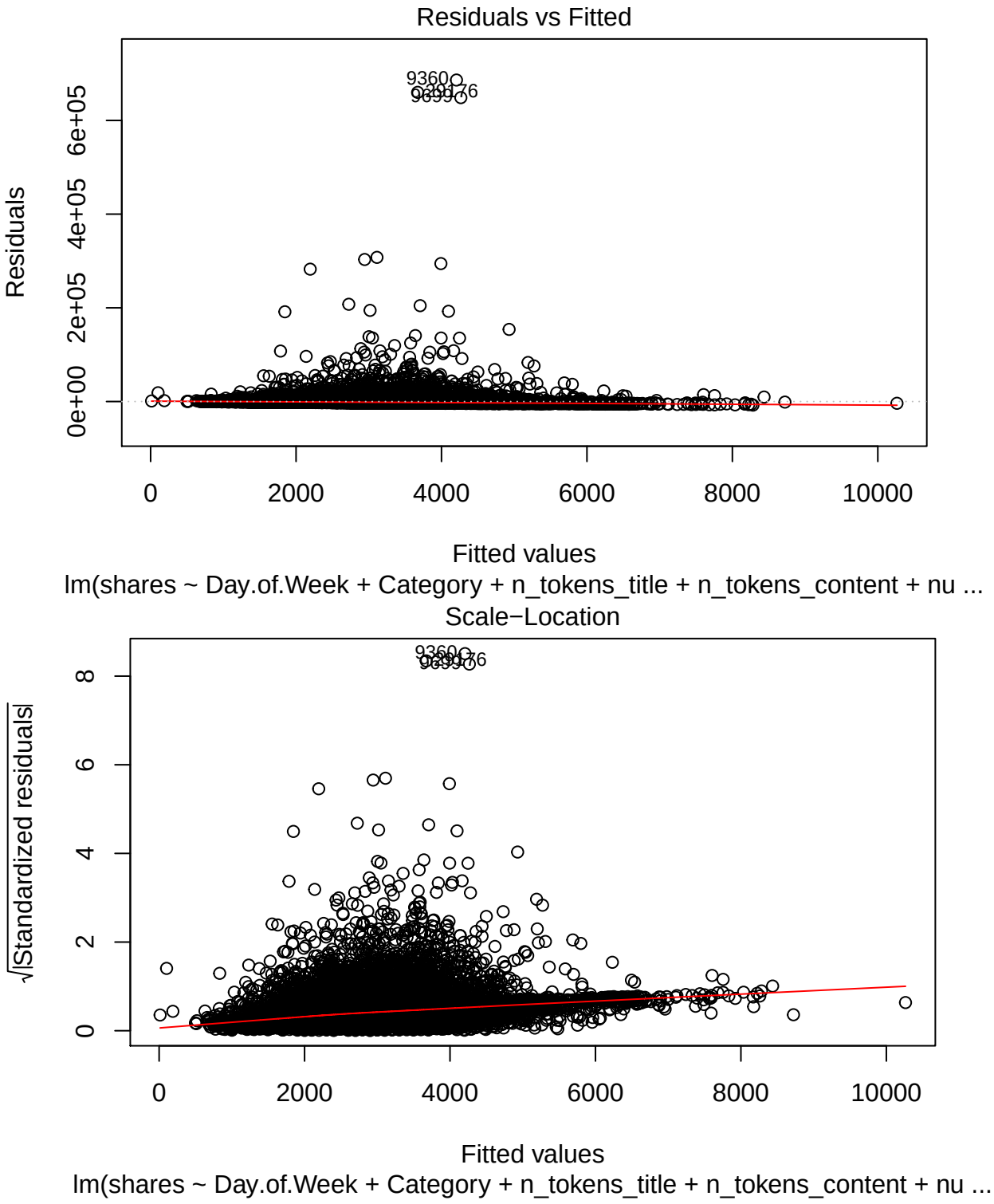


Figure x: Regression diagnostics plots for a non-transformed dependent variable

We observe the following:

- It is not clear if linearity holds because the range of residuals is very high, as seen in the fitted values vs residuals
- Constant varaince assumption does not appear to be met as seen from the scale-location plot

Here are the model plots for a log transformed variable:

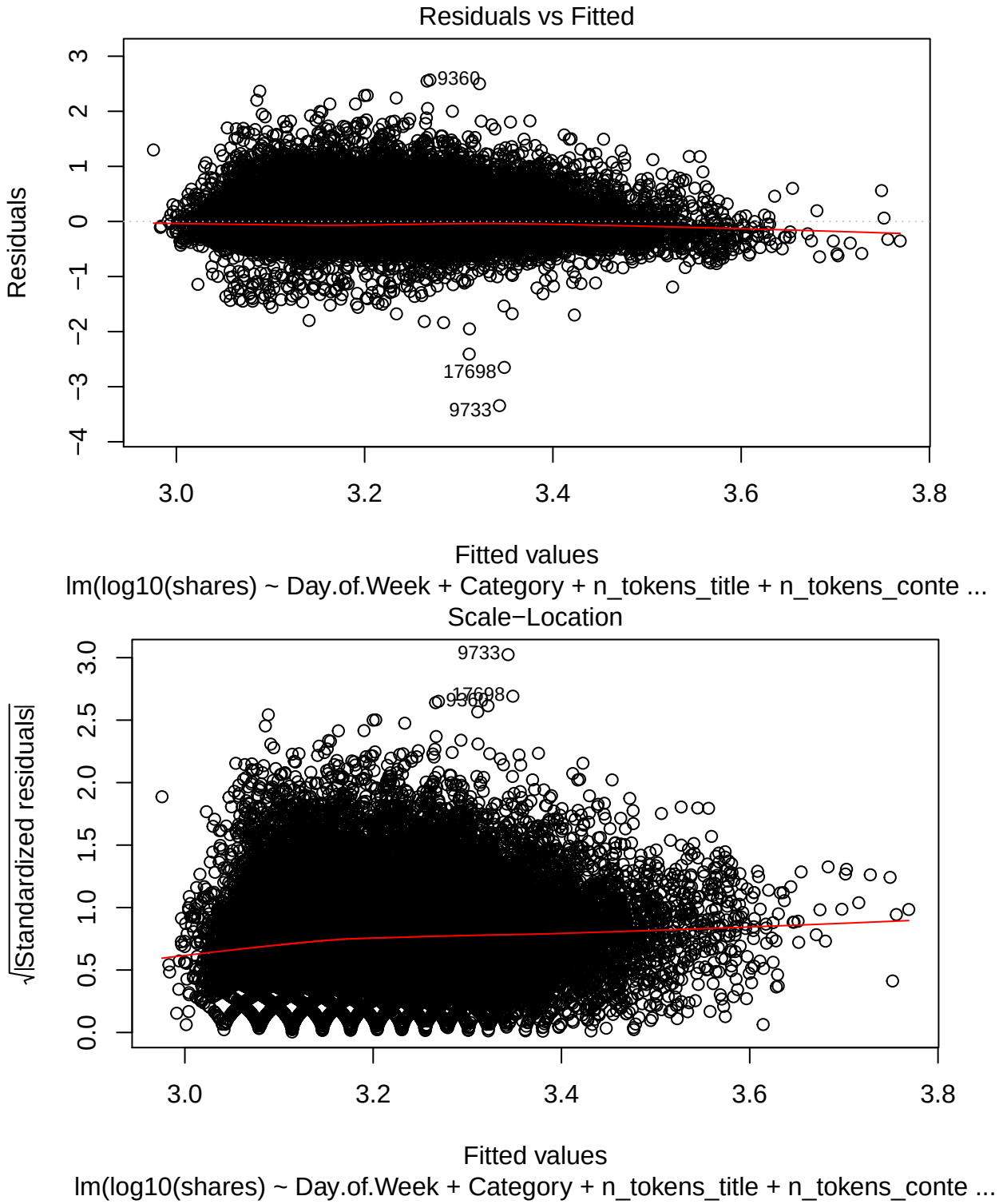


Figure x: Regression diagnostics plots for a log-transformed dependent variable

The assumptions of linearity and constant variance appear to be fulfilled for this model.

Normality assumptions