

# ADVANCED ANALYTICS

PCA, CLUSTERING & REGRESSION ASSIGNMENT

# Contents

Step 1 – Data Selection .....	2
About the data set .....	2
Choice of Dataset .....	2
Step 2 – Data Pre-Processing .....	2
Step 2.1 – Outlier Detection and Removal .....	2
Step 2.2 – Feature Selection .....	3
Step 2.3 – Variable Transformations .....	4
Step 3 – Principal Component Analysis.....	4
Step 3.1 Data Validation .....	4
Step 3.2 Calculation of Eigen Vector & Eigen Values.....	4
Step 3: Rotation of Loading Matrix & Naming PCA Components.....	5
Step 4 – Clustering Analysis .....	7
Step 4.1 – Factor Selection .....	7
Step 4.2 – Splitting of Data into Training, Test & Validation .....	7
Step 4.3 – Clustering & Validation .....	7
Training data output .....	8
Testing data output.....	8
Validation data output.....	8
Step 4.4 – Profiling of Clusters.....	9
Step 5 – Regression.....	10
Step 5.1 – Preliminary Analysis.....	10
Step 5.2 – Correlation Matrix.....	10
Step 5.3 – Model Building.....	10
First Iteration .....	10
Final Iteration.....	11
Model Equation:.....	11
Step 5.4– Evaluating Predictions.....	11

## Step 1 – Data Selection

The objective of this project is to analyse and understand popularity of online news articles and to predict if the number of shares of a given article is good or bad. This would be done by reducing the predictor variables in the data set to orthogonal variables using principal component analysis. A clustering exercise will then be performed to cluster the articles that display similarity. Finally, a regression model would be built to classify a given article as popular or unpopular.

### About the data set

The data belongs to a global, multi-platform media and entertainment company digital – Mashable. Mashable popularly publishes tech and entertainment related news articles, among others. A total of 39797 instances were recorded in the original dataset which includes 1 target, 58 predictive, and 2 non-predictive variables.

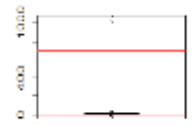
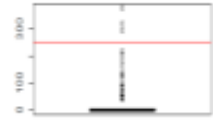


### Choice of Dataset

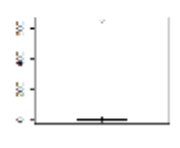
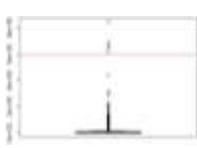
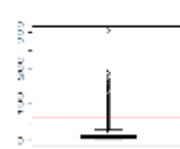

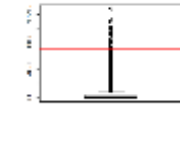
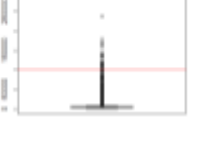

This data set provides information on the various attributes that influence the number of shares of a given news article. It was decided that this dataset would be a good choice to perform principal component analysis and clustering because it contains a large number of predictor variables. Clustering as a technique is best suited for segmentation of the target audiences or the product. Considering the current dataset, it would help to understand the underlying nature of an article and predict its popularity.

## Step 2 – Data Pre-Processing

### Step 2.1 – Outlier Detection and Removal

Extreme outliers have been removed based on the value greater than three standard deviations from mean. To ensure that too many entries were not lost, exploratory analysis has been carried out to remove the values that were below first quartile or above third quartile. Along with the statistical analyses, a business perspective was also taken into consideration while data cleaning. For example, articles containing too many links (>60) or images (>70) or videos (>60) would be very uncommon. Hence, such data entries have been removed from the data set.

Variable	Removal Criteria	Boxplot	Variable	Removal Criteria	Boxplot
n_unique_token	> 1 (as it is defining rate, values should lie between 0 -1)		kw_min_min	>250 (Quantile distance from the mean)	
n_non_stop_words	>1 (as it is defining rate, values should lie between 0 -1)		kw_avg_min	>20000 (Quantile distance from the mean)	

n_non_stop_unique_tokens	> 1 (as it is defining rate, values should lie between 0 -1)		kw_min_max	>600000 (Quantile distance from the mean)	
num_hrefs	> 60 (Business sense, uncommon in news articles)		kw_max_max	<37000 (bottom 1%)	
num_imgs	> 70 (Business sense, uncommon in news articles)		kw_max_avg	> 1,00,000 (top 1%)	
num_videos	>60 (Business sense, uncommon in news articles)				

## Step 2.2 – Feature Selection

Consider the following table for types of variable present in the dataset.

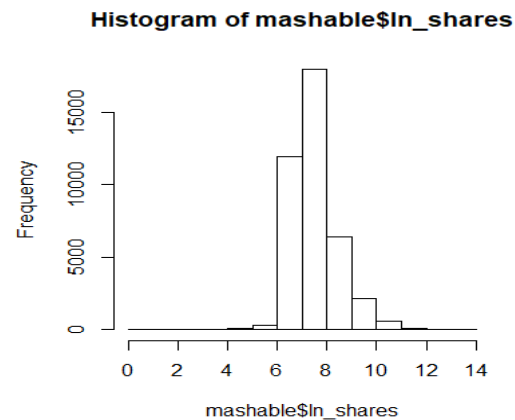
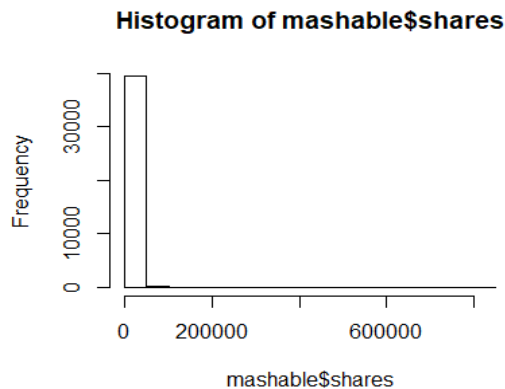
Variable Type	Count
Non- Predictive Fields (URL & Time Delta)	2
Numerical Fields	43
Categorical Field	15
Target Field	1

Categorical variables & Non- predictive field in the data set are not considered for PCA. Few numerical variables that were **not** considered for PCA as summarised below:

Variable	Meaning	Criteria for not considering in model
LDA0, LDA1, LDA2, LDA3	Closeness LDA topic 0,1,2,3	LDA (Latent Dirichlet Allocation) uses natural language processing to explain the latent clusters present in the data. So, these derived fields are not being considered.
kw_min_min, kw_max_min	Worst keyword (min/max shares)	These characteristics are represented by three Measures: Minimum, Maximum & Average. As these variables provide data about the same factor, only the Average values have been considered to provide the real representation of other attributes present in data.
kw_min_max, kw_max_max	Best keyword (min/max shares)	
kw_min_avg, kw_max_avg	Avg. keyword (min/max shares)	
self_reference_min_shares, self_reference_max_shares	Min/Max shares of referenced articles in Mashable	
min_positive_polarity, max_positive_polarity	Min/Max polarity of positive words	
min_negative_polarity, max_negative_polarity	Min/Max polarity of negative words	

### Step 2.3 – Variable Transformations

The procedure of twostep cluster analysis uses the log-likelihood distance measure assumes normal distribution for continues variables and multinomial distribution for categorical variables. The target variables “Shares” is highly skewed. A new variable “ln\_shares” is created by taking natural log of the shares variable. Also, a new target field “Popular/Unpopular” is created for the Logistic regression. Articles with count >1400 is considered as Popular and articles with count <1400 is considered as unpopular.



## Step 3 – Principal Component Analysis

### Step 3.1 Data Validation

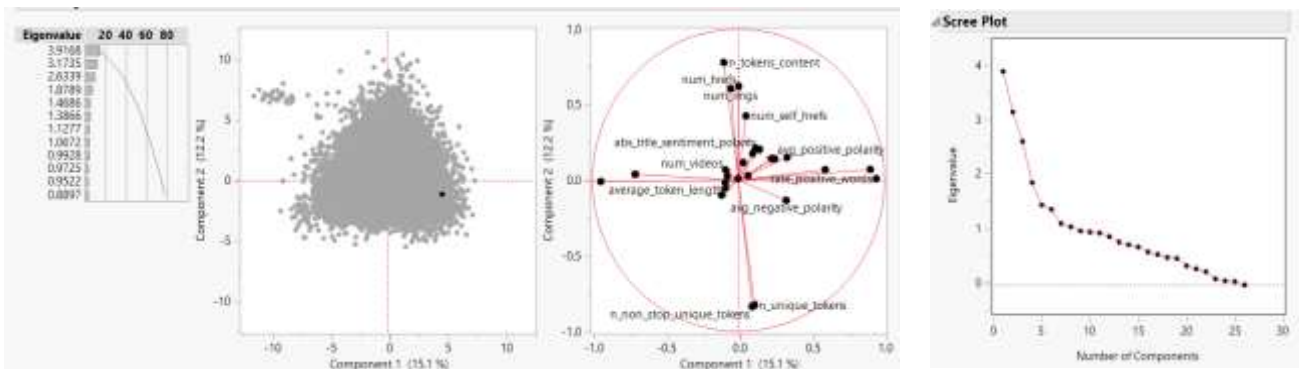
The correlation between the selected variables is in the range of 0.002 to 0.683 indicating certain variables are related, further indicating a dimension reduction requirement. Kaiser-Meyer-Olkin (KMO) of sampling adequacy and Bartlett’s test of sphericity have been thus been conducted to check whether PCA is feasible. Kaiser (1974) recommends accepting values greater than 0.5. Current dataset has KMO value of 0.6437 and Bartlett’s test gives p value as 2.2e-16. Since both these values are highly significant ( $p < 0.001$ ), PCA can be conducted on the dataset.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy. 0.618		0.6437
Bartlett's Test of Sphericity	Approx. Chi-Square	20669000
	DF	26
	p Value	<2.2e-16

### Step 3.2 Calculation of Eigen Vector & Eigen Values

A principal component analysis was conducted using JMP software on 27 input variables which resulted in the creation of 26 principal components. The PCA was performed after standardising each variable to have a mean zero and a standard deviation of one. As seen in the table and scree plot below, the first 9 principal components were found to be significant, with an eigen value of greater than or equal to one. These 9 principal components explained 67.8% of the total variance.

Component	Eigen Values			Extracted Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.9168	15.065	15.065	3.9168	15.065	15.065
2	3.1735	12.206	27.271	3.1735	12.206	27.271
3	2.6339	10.13	37.401	2.6339	10.13	37.401
4	1.8789	7.226	44.627	1.8789	7.226	44.627
5	1.4686	5.648	50.275	1.4686	5.648	50.275
6	1.3866	5.333	55.608	1.3866	5.333	55.608
7	1.1277	4.337	59.945	1.1277	4.337	59.945
8	1.0672	4.105	64.05	1.0672	4.105	64.05
9	0.9928	3.818	67.868	0.9928	3.818	67.868
10	0.9725	3.741	71.609			
11	0.9522	3.662	75.271			
12	0.8897	3.422	78.693			
13	0.789	3.034	81.727			
14	0.7394	2.844	84.571			
15	0.698	2.685	87.256			
16	0.6064	2.332	89.588			
17	0.5627	2.164	91.752			
18	0.504	1.931	93.683			
19	0.4885	1.879	95.562			
20	0.3527	1.357	96.919			
21	0.2992	1.151	98.07			
22	0.2437	0.937	99.007			
23	0.114	0.438	99.445			
24	0.0776	0.298	99.743			
25	0.0622	0.239	99.982			
26	0.0024	0.09	100			



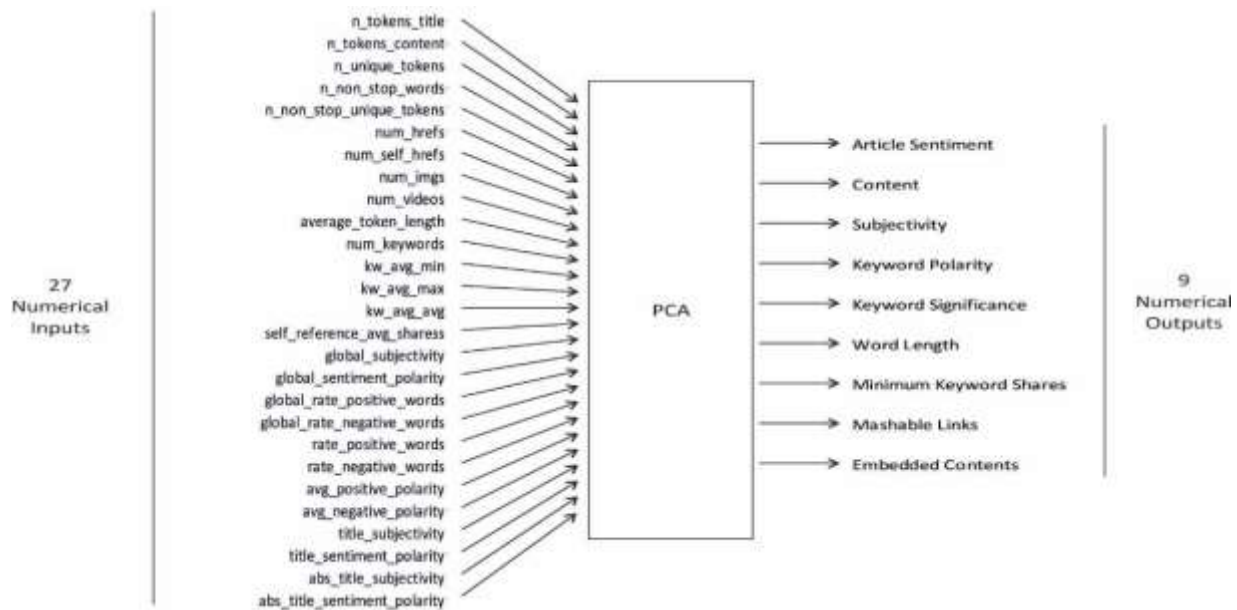
Percentage wise contribution of each Eigen Vector

Scree Plot

### Step 3: Rotation of Loading Matrix & Naming PCA Components

In unrotated factor solution the Factor "axes" may not line up very well with the pattern of variables and the loadings may show no clear pattern. Factor axes can be rotated to more closely correspond to the variables and therefore become more meaningful. Relative relationships between variables are preserved. We performed the Varimax Rotation to under the Loadings.

	Article Sentiment	Content	Subjectivity	Keyword Polarity	Keyword Significance	Word Length	Minium Keyword Shares	Mashable Links	Embedded Content
rate_positive_words	0.944								
global_sentiment_polarity	0.904								
global_rate_positive_words	0.595								
n tokens content		0.779							
num hrefs		0.623							
num_imgs		0.607							
title_subjectivity			0.684						
abs_title_subjectivity			0.673						
global subjectivity			0.489						
global_rate_negative_words	0.709								
n unique tokens									
rate_negative words	0.944								
kw_avg_avg				0.515					
kw_avg_max					0.726				
average token length						0.596			
num_keywords					-0.518				
kw_avg_min							0.650		
num_self_hrefs								0.625	
num videos									0.409
n tokens title									-0.493
avg_positive_polarity				0.396					
self reference_avg sharess									-0.424
avg negative_polarity			0.360						
title_sentiment_polarity				0.337					
abs_title_subjectivity				0.523					
n_non_stop_unique tokens		-0.837							



- The first loading vector places exactly equal weights on **rate\_positive\_words** and **rate\_negative\_words**. This vector has significant loadings of 3 other variables:

**global\_sentiment\_polarity**, **global\_rate\_negative\_words** and **global\_rate\_positive\_words**. This component roughly corresponds to the orientation in which an emotion or a sentiment is expressed that could be positive, negative or neutral. Hence this principal component was named as **Article Sentiment**.

- The second loading vector places much of its weight on aspects such as **n\_tokens\_content**, **num\_hrefs**, **num\_imgs** and **n\_non\_stop\_unique\_tokens**. As all these correspond to the content of the article, this variable was named as **Content**.
- The third vector loads the variables: **title\_subjectivity**, **abs\_title\_subjectivity**, **global\_subjectivity** and **avg\_negative\_polarity**. This roughly corresponds to the feelings, views or beliefs portrayed by the title and content of the articles. Hence, this component was named as **Subjectivity**.
- The fourth loading vector places much of its weight on **kw\_avg\_avg** and **title\_sentiment\_polarity**. This explains the polarity in titles and keywords and hence was named as **Keyword Polarity**.
- The fifth principal component relates to the keywords as it loads the **kw\_avg\_max** and **num\_keywords** variables. This component was named as **Keyword significance** as it explains the maximum number of shares a keyword can attract and the average number of keywords used in articles.
- The sixth loading vector places its total weight on a single variable which is the **average\_token\_length** (average length of words used in the articles), it was given the name **Word Length**.
- The seventh and eight principal components place their entire loadings on **kw\_avg\_min** and **num\_self\_hrefs** respectively. The seventh component was named **Minimum keyword shares** as this corresponds to the average share rate of the worst keywords used. The eighth component was named **Mashable links** as this explains the number of self-reference links included in the articles.
- The ninth component places its weight on **num\_videos**, **n\_tokens\_title** and **self\_reference\_avg\_shares**. Since this corresponds to the embedded videos and links in the articles, it was called **Embedded Content**.

#### Step 4 – Clustering Analysis

##### Step 4.1 – Factor Selection

We utilised PCA Variables (9 in number), 15 categorical variables & 1 Target Variable for the cluster Formation.

##### Step 4.2 – Splitting of Data into Training, Test & Validation

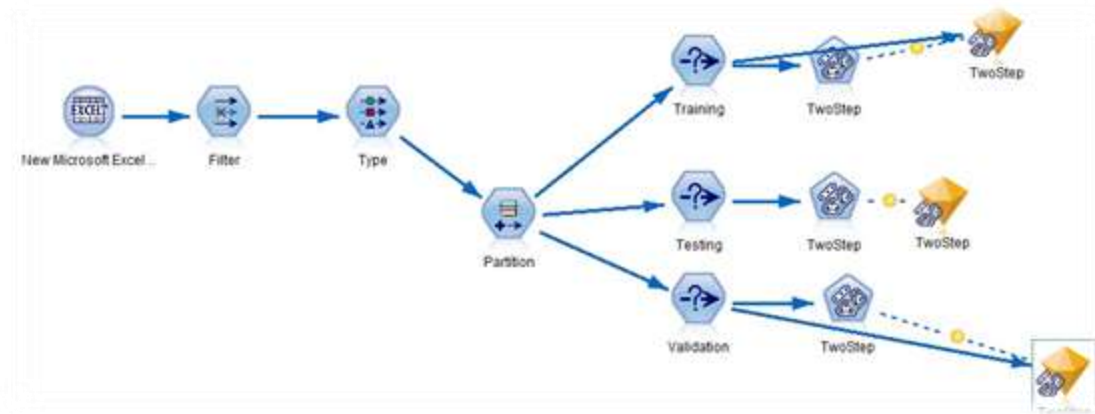
To test the relevance and structure of clusters, the data was split into 3 partitions: Training, Testing and Validation in the ratio of 50:30:20.

##### Step 4.3 – Clustering & Validation

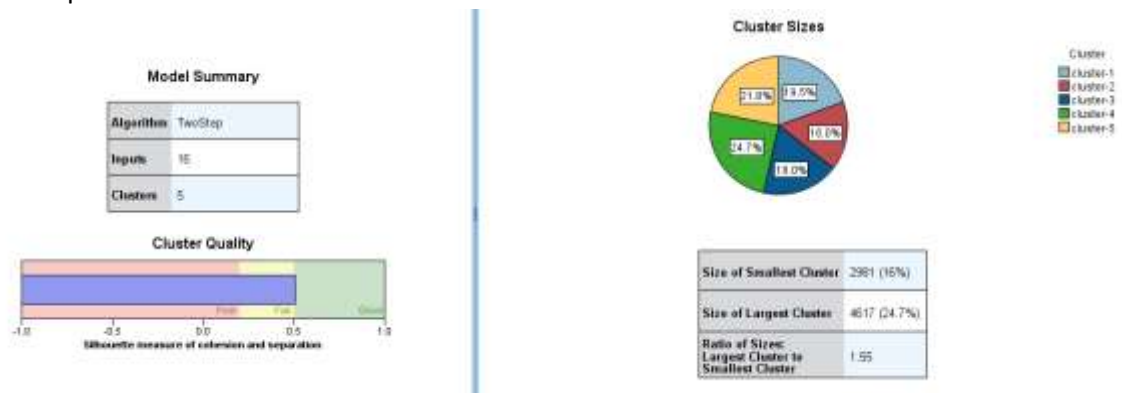
A clustering analysis was performed on SPSS Modeler using the two-step cluster method. Two step clustering algorithm was preferred over K-means because this algorithm allows inclusion of input variables that are both categorical and continuous.

After running a few iterations 9 principal components, 6 categorical variables and the predictor variable were set as input variables for the model. Dropping of 8 categorical variables resulted in an increase of silhouette ratio from 0.2 to 0.5.





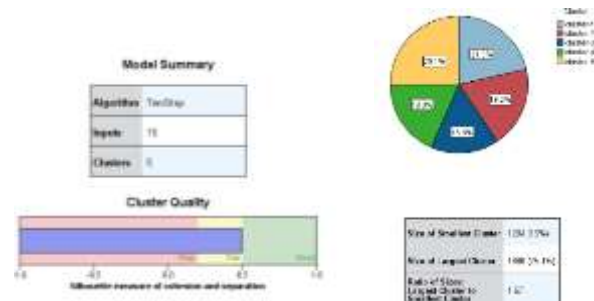
Training data output



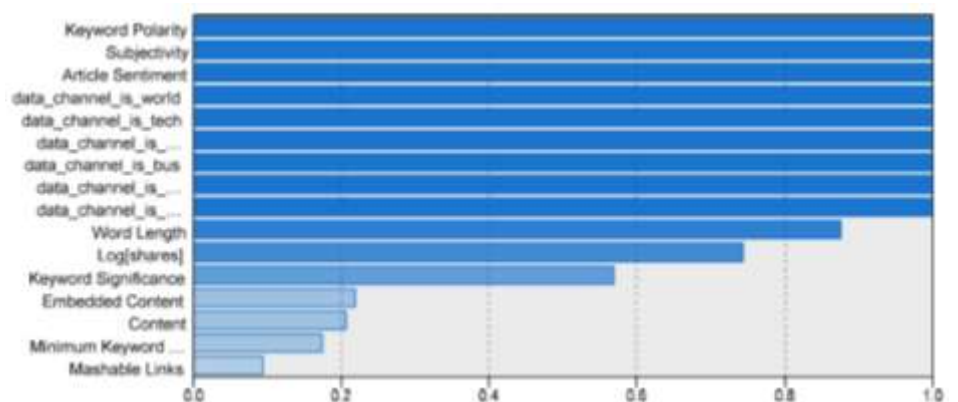
Testing data output



Validation data output



The training data output was used for interpretation and naming of clusters. The predictor importance of the input variables is as shown below:



#### Step 4.4 – Profiling of Clusters

##### **Cluster 1: Techno Articles** Cluster Size - 19%, Average share count - 3089

This cluster has articles that are high on positive sentiments with an unbiased perspective as they contain more technology related content. Although it contains comparatively few embedded videos, it has the second-best share rate.

##### **Cluster 2: Featured Biz News** Cluster Size – 16%, Average share count - 2698

This cluster has articles with more embedded videos, links and the most searched keywords. This set of articles is of interest to a target audience (business and corporate world). This explains the mediocre share rate of these articles.

##### **Cluster 3: Showbiz News:** Cluster size – 18%, Average share count - 2975

This cluster comprises of entertainment news which is both positively or negatively hyped. These articles are generally less descriptive containing short words. As this type of news is published across various digital media platforms, it captures less audience on Mashable.

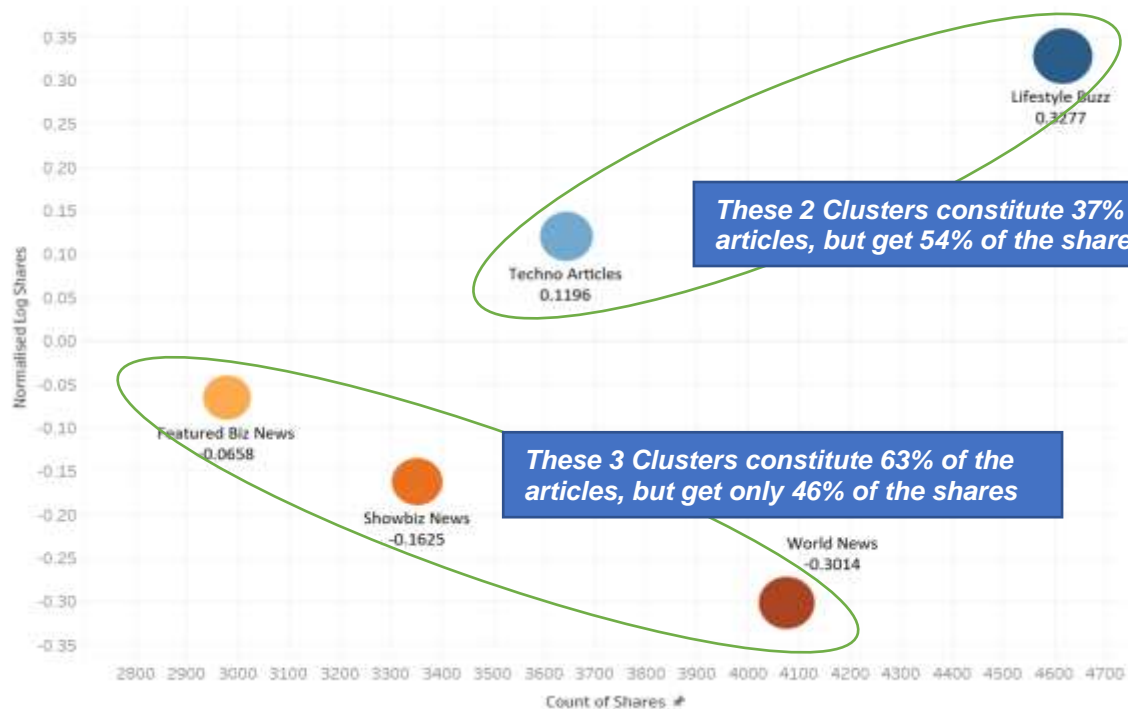
##### **Cluster 4: Lifestyle Buzz:** Cluster size - 25%, Average share count - 4595

This cluster is an amalgam of the lifestyle and social media articles as social media provides a great platform for lifestyle buzz. Hence, these articles seem to be written on a focused theme by an expert from a domain (such as travel, culinary etc.) since the subjectivity and polarity of these articles is high indicating an opinionated narration by a specialist. Also, the average word length is high demonstrating use of sophisticated domain specific vocabulary in the article. It has maximum number of shares (twice the number of shares of Cluster5) and articles that audience find insightful.

##### **Cluster 5: World News:** Cluster Size - 22%, Average share count - 2192

This cluster is a representation of world news articles which captures a very small segment of the Mashable audience as this genre is well delivered by other giant media competitors. It has a high negative sentiment value as mostly unpleasant events attract people around the globe.

Clustering of News Articles



Component	Cluster 1 Techno Article	Cluster 2 Featured Biz News	Cluster 3 Showbiz News	Cluster 4 Lifestyle Buzz	Cluster 5 World News
Article Sentiment	0.62	0.36	-0.26	0.25	-0.88
Data Channel Business	0	100	0	0	0
Data Channel Entertainment	0	0	100	0	0
Data Channel Lifestyle	0	0	0	21.2	0
Data Channel Social Media	0	0	0	22.8	0
Data Channel Technology	100	0	0	0	0
Data Channel World	0	0	0	0	100
Keyword Polarity	-0.24	-0.11	0.09	0.69	-0.59
Subjectivity	-0.45	-0.3	0.38	0.8	-0.71
Word Length	-0.15	-0.26	-0.43	0.36	0.16
Log[Shares]	0.12	-0.07	-0.16	0.33	-0.3
Keyword Significance	-0.37	0.46	-0.08	0.01	0.05
Embedded Content	-0.15	0.23	-0.06	0.05	-0.02

Table: Mean value of inputs in the cluster

## Step 5 – Regression

### Step 5.1 – Preliminary Analysis

The regression will utilise all the 9 PCA components and 15 Categorical Variables. The output variable is Popular or Unpopular

### Step 5.2 – Correlation Matrix

All PCA components are having zero correlation with each other. All Categorical Variables Chi Sq test of Independence Indicates that these are independent of each other. So, we are inputting all the 21 Variables into the model.

### Step 5.3 – Model Building

#### First Iteration

The first iterations ran with all the Principal components and the Categorical variables. Few variables p value is greater than 5% is not considered in the subsequent iteration.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6977 -1.0263 -0.7421  1.1295  1.9096

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.5478293   0.0525415   10.427 < 2e-16 ***
Article_Sentiment 0.0321217   0.0058215    5.518 3.43e-08 ***
Content         0.1175893   0.0062496   18.815 < 2e-16 ***
Subjectivity    0.1033556   0.0078029   13.246 < 2e-16 ***
Keyword_Polarity 0.1373162   0.0091467   15.013 < 2e-16 ***
Keyword_Significance 0.0039444   0.0094998    0.415 0.677993
Word_Length     0.1053676   0.0101963   10.334 < 2e-16 ***
Minimum_Keyword_Shares 0.1588022   0.0124459   12.759 < 2e-16 ***
Mashable_Links -0.0624458   0.0109628   -5.696 1.23e-08 ***
Embedded_Content -0.0002938   0.0118928   -0.025 0.980291
data_channel_is_lifestyle1 0.0058695   0.0580856    0.101 0.919512
data_channel_is_entertainment1 -0.6647704   0.0429928  -15.462 < 2e-16 ***
data_channel_is_bus1 -0.0452014   0.0450915   -1.002 0.316132
data_channel_is_socmed1  0.8699271   0.0594940   14.622 < 2e-16 ***
data_channel_is_tech1    0.3236075   0.0450271    7.187 6.63e-13 ***
data_channel_is_world1 -0.5645588   0.0465989  -12.115 < 2e-16 ***
weekday_is_monday1 -0.6588552   0.0498277  -13.223 < 2e-16 ***
weekday_is_tuesday1 -0.7645533   0.0490849  -15.576 < 2e-16 ***
weekday_is_wednesday1 -0.7649240   0.0491336  -15.568 < 2e-16 ***
weekday_is_thursday1 -0.7012237   0.0491998  -14.253 < 2e-16 ***
weekday_is_friday1 -0.5514893   0.0507832  -10.860 < 2e-16 ***
weekday_is_saturday1  0.2216492   0.0624152    3.551 0.000383 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 51554  on 37388  degrees of freedom
Residual deviance: 47684  on 37367  degrees of freedom
AIC: 47728
```

## Final Iteration

P values of all the variables inputs is significant and it represent the final model.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7336 -1.0285 -0.7411  1.1308  1.9140

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.527375   0.044607  11.823 < 2e-16 ***
Article_Sentiment  0.032099   0.005813   5.522 3.35e-08 ***
Content      0.118394   0.006214  19.052 < 2e-16 ***
Subjectivity  0.105884   0.007206  14.694 < 2e-16 ***
Keyword_Polarity  0.140113   0.008572  16.345 < 2e-16 ***
Word_Length   0.108016   0.009902  10.909 < 2e-16 ***
Minimum_Keyword_Shares  0.160037   0.012367  12.941 < 2e-16 ***
Mashable_Links -0.062222   0.010920  -5.698 1.21e-08 ***
data_channel_is_entertainment1 -0.644418   0.032797 -19.648 < 2e-16 ***
data_channel_is_socmed1      0.891695   0.051290  17.386 < 2e-16 ***
data_channel_is_tech1       0.345745   0.031763  10.885 < 2e-16 ***
data_channel_is_world1     -0.540039   0.033854 -15.952 < 2e-16 ***
weekday_is_monday1         -0.660393   0.049769 -13.269 < 2e-16 ***
weekday_is_tuesday1        -0.765444   0.049041 -15.608 < 2e-16 ***
weekday_is_wednesday1      -0.766018   0.049090 -15.604 < 2e-16 ***
weekday_is_thursday1       -0.702416   0.049146 -14.292 < 2e-16 ***
weekday_is_friday1         -0.551979   0.050762 -10.874 < 2e-16 ***
weekday_is_saturday1       0.221909   0.062398   3.556 0.000376 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 51554  on 37388  degrees of freedom
Residual deviance: 47685  on 37371  degrees of freedom
AIC: 47721
```

## Model Equation:

$$\log\left(\frac{p}{1-p}\right) = -0.52 + 0.032 * Article_{Sentiment} + 0.11 * Content + 0.10 * Subjectivity + 0.14 * Keyword\_Polarity + 0.10 * Word\_Length + 0.16 * Minimum_{Keyword\_Share} - 0.06 * Mashable_{Links} - 0.64 * datachannel_{Entertainment} + 0.89 * datachannel_{Socialmedia} + 0.34 * datachannel_{Technology} - 0.54 * datachannel_{World} - 0.66 * weekday_{monday} - 0.77 * weekday_{tuesday} - 0.77 * weekday_{wednesday} - 0.70 * weekday_{thursday} - 0.55 * weekday_{friday} + 0.22 * weekday_{saturday}$$

## Step 5.4– Evaluating Predictions

The estimated model (for cut-off = 0.5) correctly predicts the true popular articles 61.62% of the time and the true non-popular articles 66.94% of the time. Accuracy is 64.14%

```
> confusionMatrix(p_class,news$Article, positive="1")
Confusion Matrix and Statistics

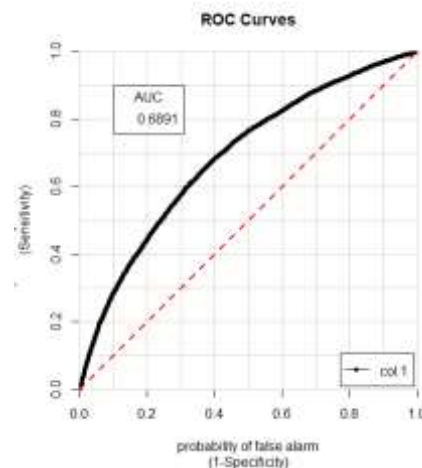
          Reference
Prediction  0      1
 0  12754  7110
 1   6299 11226

      Accuracy : 0.6414
      95% CI   : (0.6365, 0.6462)
 No Information Rate : 0.5096
 P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.2819
McNemar's Test P-Value : 2.653e-12

      Sensitivity : 0.6122
      Specificity : 0.6694
   Pos Pred Value : 0.6406
   Neg Pred Value : 0.6421
      Prevalence : 0.4904
   Detection Rate : 0.3002
   Detection Prevalence : 0.4687
   Balanced Accuracy : 0.6408

'Positive' Class : 1
```



The Sensitivity of the model can be improved by exploring other factors, such time of release of the article, web interaction data for the usage, the various social media responses & the time series popularity data of Mashable. Due to the presence of interactions and higher order terms of the main effects, interpreting the shares and popularity of article is complex.