**Project Report on Social marketing- analyst**

Name:

**Project: -** We have perform here social-marketing analysis on given data. Dataset has 39000 articles with the huge-number (with the collective of 31) of highpoint were separated from the HTML code of the article, it includes the titles and the constituents of each article. Here the channel classes are: 'Internet based life', 'Way of life', 'Business', 'Diversion', 'Innovation', and 'World'. Furthermore, a few natural language treatment highpoint were furthermore unraveled.

We have open and explore the given dataset to understand & manage the all six type of data channel:( socmed, tech, world ,lifestyle, entertainment, bus,) and the connecting data. In all 3 data channel, the value 1 represent that the data the row is of the consistent data channels.

Then we have copied the separate dataset for all channel to dissimilar Excel sheet ( to sort and filter )by all data channel.

Then we have Imported the online popularity data file by using read\_xlsx() function on R studio. Then we have installed and loaded the required libraries for this project.

* **Explanatory0DataoAnalysis:**

Data set has 39644 row & 31 column. It has 356 missing data NA.

By help of Mics package we found top five lower and top five the highest, unique, mean, percentiles, missing values also etc.. It show all the explanatory data-analysis(EDA):

**By using Misc package describe function has given this output:**

31 Variables 39644 Observations

------------------------------------------------------------------------------------------------------------

shares

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 1454 0.999 3395 4149 584 708 946 1400 2800 6200

.95

10800

lowest : 1 4 5 8 22, highest: 617900 652900 663600 690400 843300

------------------------------------------------------------------------------------------------------------

timedelta

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 724 1 354.5 247 43 71 164 339 542 661

.95

697

lowest : 8 9 10 11 12, highest: 727 728 729 730 731

------------------------------------------------------------------------------------------------------------

n\_tokens\_title

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 20 0.98 10.4 2.365 7 8 9 10 12 13

.95

14

Value 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Frequency 1 5 28 184 805 2120 4227 6278 7358 6960 5299 3509 1733 756 259 91

Proportion 0.000 0.000 0.001 0.005 0.020 0.053 0.107 0.158 0.186 0.176 0.134 0.089 0.044 0.019 0.007 0.002

Value 18 19 20 23

Frequency 22 6 2 1

Proportion 0.001 0.000 0.000 0.000

------------------------------------------------------------------------------------------------------------

n\_tokens\_content

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 2406 1 546.5 452.1 105 152 246 409 716 1090

.95

1407

lowest : 0 18 21 22 24, highest: 7081 7185 7413 7764 8474

------------------------------------------------------------------------------------------------------------

n\_unique\_tokens

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 24878 1 0.5482 0.1747 0.3491 0.4064 0.4709 0.5392 0.6087 0.6767

.95

0.7209

Value 0 700

Frequency 39643 1

Proportion 1 0

------------------------------------------------------------------------------------------------------------

n\_non\_stop\_words

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 52 0.985 0.9965 0.1103 1 1 1 1 1 1

.95

1

Value 0 1042

Frequency 39643 1

Proportion 1 0

------------------------------------------------------------------------------------------------------------

n\_non\_stop\_unique\_tokens

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 21253 1 0.6892 0.1775 0.4773 0.5534 0.6257 0.6905 0.7546 0.8188

.95

0.8571

Value 0 650

Frequency 39643 1

Proportion 1 0

------------------------------------------------------------------------------------------------------------

num\_hrefs

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 133 0.997 10.88 10.04 1 2 4 8 14 23

.95

30

lowest : 0 1 2 3 4, highest: 162 171 186 187 304

------------------------------------------------------------------------------------------------------------

num\_self\_hrefs

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 59 0.979 3.294 3.216 0 0 1 3 4 6

.95

9

lowest : 0 1 2 3 4, highest: 62 63 65 74 116

------------------------------------------------------------------------------------------------------------

num\_imgs

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 91 0.899 4.544 6.388 0 0 1 1 4 14

.95

20

lowest : 0 1 2 3 4, highest: 100 101 108 111 128

------------------------------------------------------------------------------------------------------------

num\_videos

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 53 0.735 1.25 2.136 0 0 0 0 1 2

.95

6

lowest : 0 1 2 3 4, highest: 66 73 74 75 91

------------------------------------------------------------------------------------------------------------

average\_token\_length

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 30136 1 4.548 0.5657 4.156 4.303 4.478 4.664 4.855 5.037

.95

5.153

lowest : 0.000000 3.600000 3.624585 3.653846 3.657143, highest: 6.816754 7.218430 7.695652 7.974684 8.041534

------------------------------------------------------------------------------------------------------------

num\_keywords

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90

39644 0 10 0.976 7.224 2.165 4 5 6 7 9 10

.95

10

Value 1 2 3 4 5 6 7 8 9 10

Frequency 51 45 635 2427 4829 6801 7322 6094 4732 6708

Proportion 0.001 0.001 0.016 0.061 0.122 0.172 0.185 0.154 0.119 0.169

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_lifestyle

n missing distinct Info Sum Mean Gmd

39644 0 2 0.15 2099 0.05295 0.1003

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_entertainment

n missing distinct Info Sum Mean Gmd

39644 0 2 0.439 7057 0.178 0.2927

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_bus

n missing distinct Info Sum Mean Gmd

39644 0 2 0.399 6258 0.1579 0.2659

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_socmed

n missing distinct Info Sum Mean Gmd

39644 0 2 0.165 2323 0.0586 0.1103

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_tech

n missing distinct Info Sum Mean Gmd

39644 0 2 0.453 7346 0.1853 0.3019

------------------------------------------------------------------------------------------------------------

data\_channel\_is\_world

n missing distinct Info Sum Mean Gmd

39644 0 2 0.502 8427 0.2126 0.3348

------------------------------------------------------------------------------------------------------------

weekday\_is\_monday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.419 6661 0.168 0.2796

------------------------------------------------------------------------------------------------------------

weekday\_is\_tuesday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.455 7390 0.1864 0.3033

------------------------------------------------------------------------------------------------------------

weekday\_is\_wednesday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.457 7435 0.1875 0.3048

------------------------------------------------------------------------------------------------------------

weekday\_is\_thursday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.449 7267 0.1833 0.2994

------------------------------------------------------------------------------------------------------------

weekday\_is\_friday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.369 5701 0.1438 0.2463

------------------------------------------------------------------------------------------------------------

weekday\_is\_saturday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.174 2453 0.06188 0.1161

------------------------------------------------------------------------------------------------------------

weekday\_is\_sunday

n missing distinct Info Sum Mean Gmd

39644 0 2 0.193 2737 0.06904 0.1285

------------------------------------------------------------------------------------------------------------

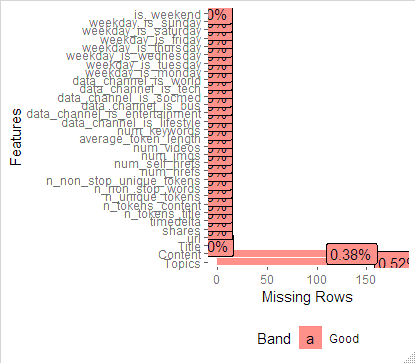
is\_weekend

n missing distinct Info Sum Mean Gmd

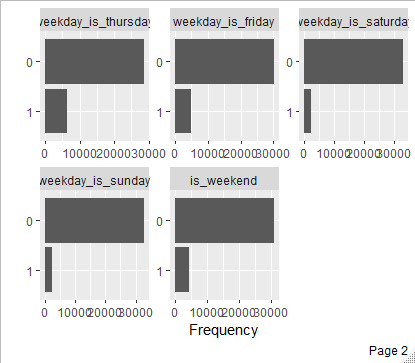
39644 0 2 0.341 5190 0.1309 0.2276

------------------------------------------------------------------------------------------------------------

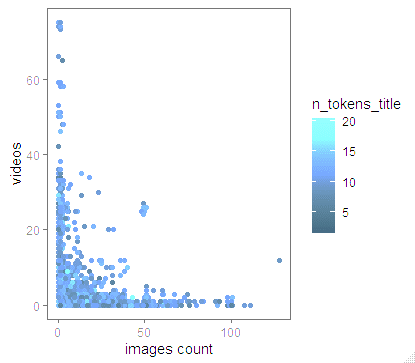
**Missing0values in data0set plot:**



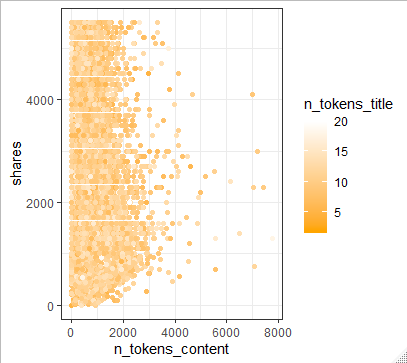
**Frequency0of Variables0for weekday0and weekend:**



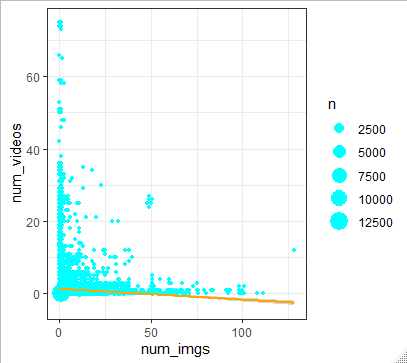
**Plot0for Count of images and0videos:**



**Plot0for n\_tokens\_content0with0shares:**



**Plotting0for Plot the x=num\_imgs,0y=num\_videos:-**



**0Fit LM( linear0model) for0shares , n\_tokens0content**

**onlinel = lm(shares ~ n\_tokens\_content,**

**data = onlineout)**

**summary(onlinel):**

**0Output**:

0Call:

lm(formula = shares ~ n\_tokens\_content, data = onlineout)

**0Residuals:**

Min 1Q Median 3Q Max

-1724.1 -767.1 -368.1 425.7 3890.4

0Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.610e+03 9.105e+00 176.79 <2e-16 \*\*\*

n\_tokens\_content 1.140e-01 1.271e-02 8.97 <2e-16 \*\*\*

---

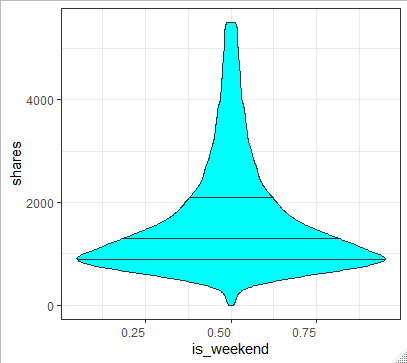
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1102 on 35101 degrees of freedom

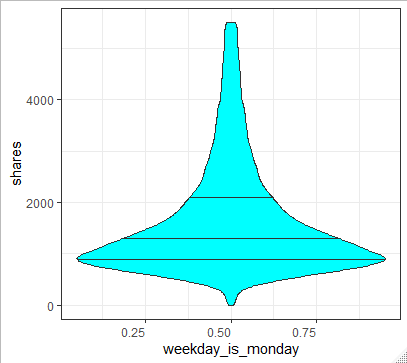
Multiple R-squared: 0.002287, Adjusted R-squared: 0.002258

F-statistic: 80.45 on 1 and 35101 DF, p-value: < 2.2e-16

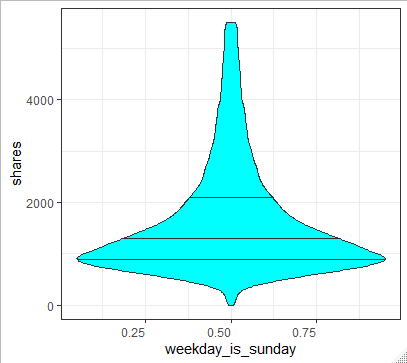
**Plot0for the is0weekend with0Shares:**



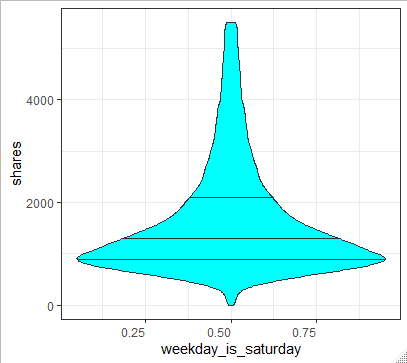
**Plot0of weekday is0monday vs0Shares:**



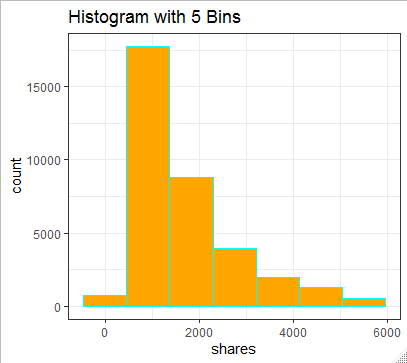
**Plot0of weekday0is sunday with0Shares:**

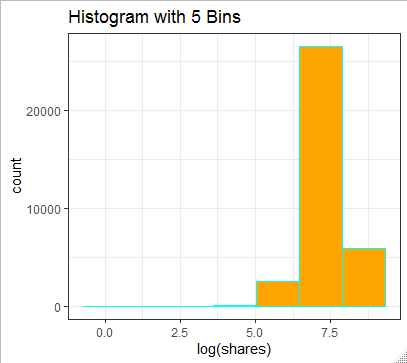


**Plot0for weekday0is saturday with0Shares:-**



**Numbers of0cluster:-**





We have used dataset to comprehend the given types of data channel and the0relating data.0Each data channel0column, here the0value of 1 represent that0data in 0row is of the0consistent data0 channel.

**Data0channel is0lifestyle:-**

0 1

33305 1798

**Data0channel is0entertainment:-**

0 1

28756 6347

**Data0channel is0bus:-**

0 1

29354 5749

**Data0channel is0socmed:-**

0 1

33114 1989

**Data0channel is0tech:-**

0 1

28554 6549

**Data0channel is0world:-**

0 1

27225 7878

**0Correlation b/w in numeric0variables:-**

corrln<-cor(dat num[,c(2:25)])

Install the packages for correlation

install.packages("corrgram")

loading the packages for correlation

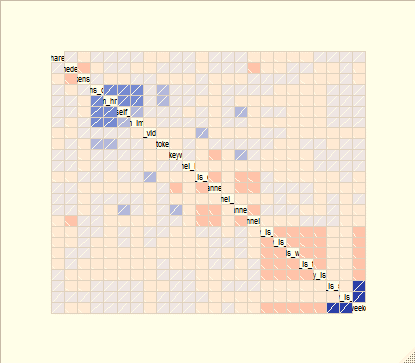
library(corrgram)

corma<-corrgram(corrln)

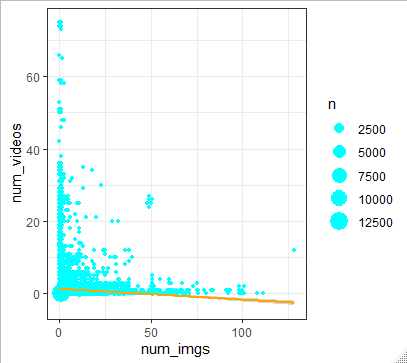
saved the file

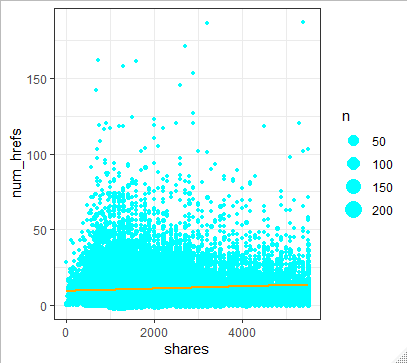
write.csv(corma,"Correlation.csv")

**Output:**



**Fitted0line to our0plot to examine0the correlation0for num imgs0and num videos0continuous variables-:**





**References**

1. Alam, M. H., Ryu, W. & Lee, S. (2016) ‘Joint multi-grain topic sentiment: modeling semantic aspects for online reviews.’ Information Sciences [Online]. 339: 206-223.
2. Fuggle, L. (2017) 70 travel & tourism statistics to know about in 2016. Available from: https://www.trekksoft.com/en/blog/travel-tourism-stats-2016 [Accessed on 15 March 2017]. Gémar, G. & Jiménez-Quintero, J. A. (2015) ‘Text mining social media for competitive analysis.’ Tourism & Management Studies [Online]. 11 (1): 84-90.
3. Guo, Y., Barnes, S. & Jia, Q. (2017) ‘Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation.’ Tourism Management [Online]. 59: 467-483.
4. He, W., Zha, S. & Li, L. (2013) ‘Social media competitive analysis and text mining: A case study in the pizza industry.’ International Journal of Information Management .

.

**ONLINE POPULARITY OF DATA CHANNELS**

[Author Name(s), First M. Last, Omit Titles and Degrees]

**Sentiment analysis on Housing tweets:** It is a critical development for giving bits of learning to update the business essential worry in campaign following, customer driven advancing system and brand care. End assessment techniques are used to convey feeling classes for instance, 'positive', 'negative' and 'unprejudiced'.

**Explanatory text analysis:** we have given the housing tweets to perform text analysis and perform the sentiment analysis. We have given two sentiment positive and negative tweets of two columns and 263 tweets.

# ‘get\_nrc\_sentiment’: I have Developed a dictionary-based sentiment analytics engine based on R library ‘”syuzhet”’ to perform the analysis the different emotions from hotel review tweets .

I have analysis and aggregate the eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) from the hotel review tweets data file ‘hotel\_tweets.csv’ by using the function **‘get\_nrc\_sentiment’**.

**Using the function 'get\_nrc\_sentiment'**: It carryout the sentiment mining using the get\_nrc\_sentiment () function log the findings under a variable result:

Code:-result <- get\_nrc\_sentiment(as.character(mydataCopy)).

First I have perform the analysis and text mining on negative tweets to Analyze and aggregate the eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust). By using plot we can see all these eight emotions:

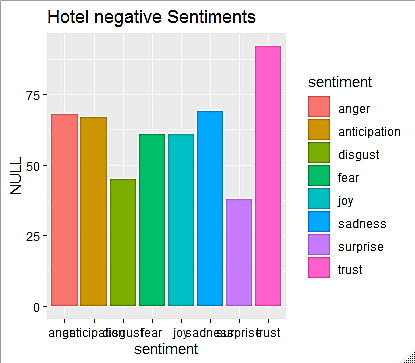


Fig: the hotel negative tweets sentiment analysis with emotions

**Text mining on Positive tweets**

Now I have again execute the analysis and text mining  **positive tweets** to Analyze and aggregate the eights emotion (**anger, anticipation, disgust, fear, joy, sadness, surprise and trust**). By using bar plot we can see all these eight emotions:

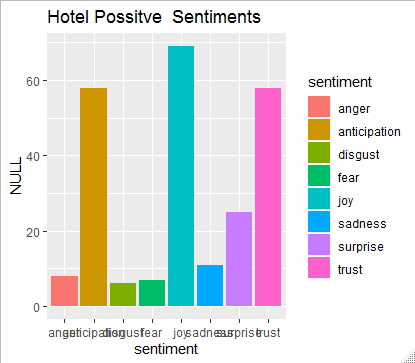


Fig: hotel positive tweets sentiment analysis with emotions



Figure: The hotel positive tweets sentiment analysis with emotions

In the positive tweets column of data set have some negative tweets sentiment.

**Q3:We Developed machine learning based model using the R libraries**

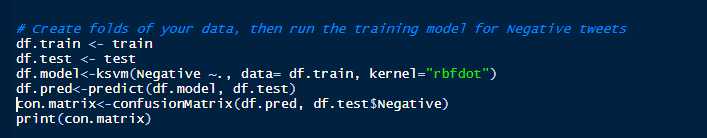
'tm' and 'e1071'

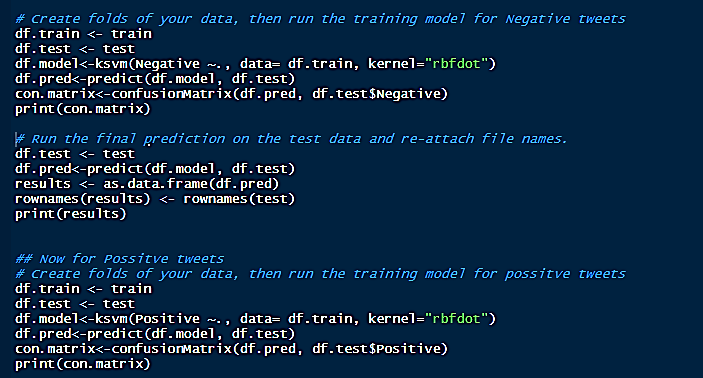
We have evaluate the predictive precisions by the SVM classifier

**We have performed the Knn Classifie for the text classification by using Knn**

df.model<-ksvm(Negative ~., data= df.train, kernel="rbfdot")

We evaluated the testing accuracies and report the predicted result.





References

1. Kanimozhi KV, Venkatesan M. Unstructured data analysis-a survey. Int J Adv Res Comput Commun Eng. 2015;4:223–5.[CrossRef](https://doi.org/10.17148/IJARCCE.2015.4354)[Google Scholar](http://scholar.google.com/scholar_lookup?title=Unstructured%20data%20analysis-a%20survey&author=KV.%20Kanimozhi&author=M.%20Venkatesan&journal=Int%20J%20Adv%20Res%20Comput%20Commun%20Eng&volume=4&pages=223-225&publication_year=2015)
2. Grar M, Cherepnalkoski D, Mozeti I, Kralj Novak P. Stance and influence of twitter users regarding the brexit referendum. Comput Social Netw. 2017;4:6.[CrossRef](https://doi.org/10.1186/s40649-017-0042-6)[Google Scholar](http://scholar.google.com/scholar_lookup?title=Stance%20and%20influence%20of%20twitter%20users%20regarding%20the%20brexit%20referendum&author=M.%20Grar&author=D.%20Cherepnalkoski&author=I.%20Mozeti&author=P.%20Kralj%C2%A0Novak&journal=Comput%20Social%20Netw&volume=4&pages=6&publication_year=2017)
3. Fan W, Wallace L, Rich S, Zhang Z. Tapping the power of text mining. Commun ACM. 2006;49(9):76–82.[CrossRef](https://doi.org/10.1145/1151030.1151032)[Google Scholar](http://scholar.google.com/scholar_lookup?title=Tapping%20the%20power%20of%20text%20mining&author=W.%20Fan&author=L.%20Wallace&author=S.%20Rich&author=Z.%20Zhang&journal=Commun%20ACM&volume=49&issue=9&pages=76-82&publication_year=2006)

.