Practical Data Science Course Project:

Digital Social Media: The Secrets Behind Facebook's Posts

Wenyuan Lu

University of Pittsburgh
Course: BMIS 2542 Programming Essentials for Data Science
Professor: Narayan Ramasubbu

Digital Social Media: The Secrets Behind Facebook's Posts

1 Introduction

1.1 Project Background

The Web has become a more social environment since its beginning. There has never before existed such an environment that has the ability to link a piece of content to another. The Web creates content webs that contain value. It has been observed that traditional media has fallen to the side since the widespread use of social media such as blogs, Twitter, Facebook.

From the marketing side, social media and social media applications that build consumer communities involving rich user-generated content are new marketplaces and or tools for marketers These 'status updates' short posts by Facebook users to update their online friends on what their current state of affairs or emotions are, are examples of how pervasive social networking sites(SNSs) such as Facebook can be. Not only for the individual users, for business, the page is the window people may look through. The 'status updates', including photos, status, videos are the ways that business interact with their 'follower's.

To evaluate the relationships between the post contents and the number of reactions, the number of shares, the number of comments, to help the businesses to run their Facebook pages more efficiently and economically, I took car industry and cosmetic industry as examples. For each industry, I randomly chose 10 companies. Meanwhile, I also chose 10 celebrities to see the difference between the companies' pages and the personals' pages on Facebook.

1.2 Project Target

The main purpose of this project is to research the relations between the post content organizations posted on Facebook 'page's, and the numbers of likes, shares, comments and ultimately generate some practical suggestions for people who is running digital social media and who are trying to spread their posts and get more reactions.

"Make the users talk themselves" and to give suggestions are the aims of this project.

2 Literature Review

2.1 Overview of Digital Social Media Marketing

Social media marketing can be simply defined as the use of social media channels to promote a company and its products. This kind of marketing can be thought of as a subset of online marketing activities that complete traditional Web-based promotion strategies, such as e mail newsletters and online advertising campaigns. Traditional marketing is brand generated. The content is completely from brand to the customer. On the contrary, social media marketing is comprised of new features, such as the following:

1) Web-based social networking promoting comprises of multidirectional discoursed. Brands converse with the clients, clients converse with the brands, and—perhaps above all—clients converse with each other. This situation is a new type of engagement that was impossible before Web 2.0.

2) Social media marketing is participatory. Web-based social networking promoting comprises of multidirectional discoursed. Brands converse with the clients, clients converse with the brands, and—perhaps above all—clients converse with each other.

3) Online networking promoting is client produced. The greater part of the substance and associations in an online group are made by clients, not by the brand. Obviously, some substance and discussions are created by the brand, however these sorts of substance and discussions are few. Online networking promoting is client produced. The greater part of the substance and associations in an online group are made by clients, not by the brand. Obviously, some substance and discussions are created by the brand, however these sorts of substance and discussions are few.

"The aim is to make users talk."

According to Figure 1 which is the social media strategy map, social media strategy should be aligned with the customer journey map, which are respectively awareness, consider, intent, purchase, frequently, loyalty, advocacy, insight. For each step, the social media strategy should be different.

Figure 1 Social media strategy map

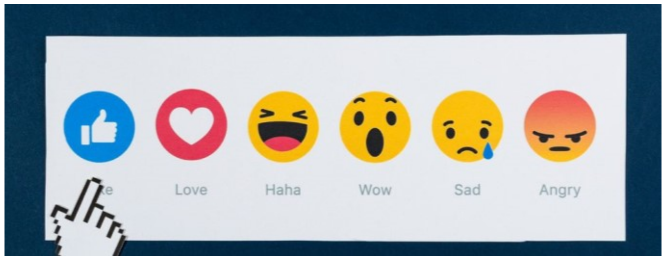
In the social media strategy map, Facebook should be seated at the first stage, for this project because this project is going to use the data from the pages of businesses. Pages, the windows for businesses to communicate with customers, may handle the responsibility that introducing the new products, branding their products, and finally help more people to know the products and understand the company cultures as well as the up-to-date news.

2.2 Facebook's 'Data'

Facebook is an American online social media and social networking service company based in Menlo Park, California. Its website was launched on February 4, 2004, by Mark Zuckerberg. In February 2015, Facebook announced that it had reached two million active advertisers with most of the gain coming from small businesses. An active advertiser is an advertiser that has advertised on the Facebook platform in the last 28 days. In March 2016, Facebook announced that it reached three million active advertisers with more than 70% from outside the US.

Launched, as a ''social utility, Facebook helps people communicate more efficiently with their friends, family and coworkers'' (Facebook Factsheet, 2010, para. 1). By building a profile, each Facebook user is able to post notes, photos, links, and videos to be shared with ''friends''; that is, other members who are connected to an individual's online social network, and thus granted access to view the individual's profile. The ''Home Page'' allows each Facebook user to be constantly updated on the most recent postings and interactions of and among friends. Facebook users can also enable 'Facebook Chat' to instant message online friends in real time.

Figure 2

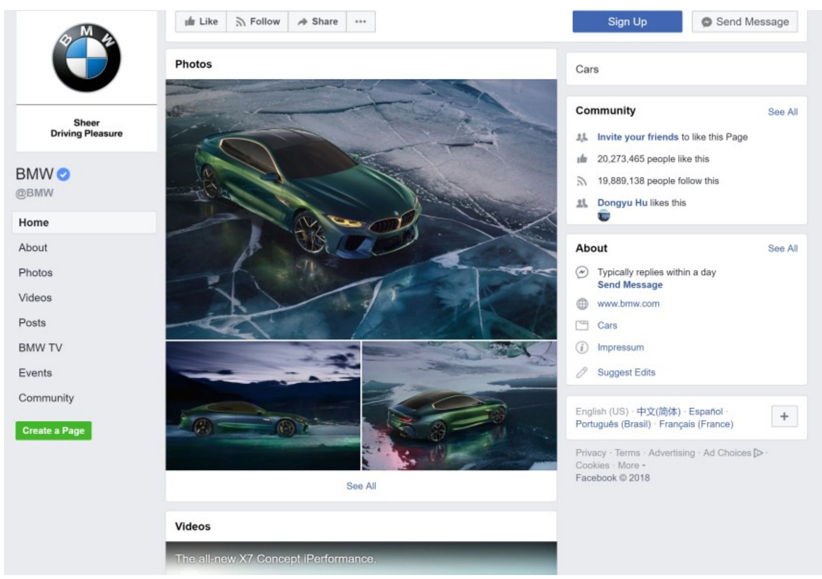


Source: https://www.theverge.com/2018/4/12/17226536/how-to-quit-facebook-tips-from-the-year-with-no-internet-guy

Currently, the users are able to express their feelings by clicking the different emojis including the like, love, haha, wow, sad and angry faces instead of 'liking' only.

Here is what a typical Facebook page looks like. The business can post what they want their followers to see including texts, images, and videos.

Figure 3



Source: https://www.facebook.com/BMW/

3 Data Collection and Exploration

3.1 Data Collection

The data was collected from Facebook's websites utilizing the API called graph API developed by Facebook with Python. I also utilize the IBM Watson Natural Language Processing API to analyze the post texts, as well as the Google Vision API to analyze the images people upload.

According to the official Facebook Graph API introduction, the Graph API is the primary way to get data into and out of the Facebook platform. It's a low-level HTTP-based API that apps can use to programmatically query data, post new stories, manage ads, upload photos, and perform a wide variety of other tasks.

IBM Watson was created as a question answering (QA) computing system that IBM built to apply advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering. The Google Vision API allows developers to easily integrate vision detection features within applications, including image labeling, face and landmark detection, optical character recognition (OCR), and tagging of explicit content.

With the help to all the APIs, I firstly collected data including the status_idstatus_link, status_published, num_reactions, num_comments, num_shares, num_likes, num_love and so on.

To have a better view of the overall pattern and rules, I choose three main categories which are car industry, cosmetic industry, and several celebrities to see the potential differences between the business pages and the personal pages. Table documents all the names of companies or celebrities I collected.

Table 1

car	cosmetic	celebrity
bmw	armanibeauty	barackobama
calillac	belifusa	brunomars
chevroelt	cliniqueus	donaldtrump
ford	karitybeauty	dwaynejohnson
gmc	lamer	jackie
honda	laprairie	justinbieber
hyundai	loccitaneusa	katyperry
jeep	maybelline	ladygaga
kia	narscosmetics	leomessi
lincoln	origins	linkinpark
mercedes-benz	sephora	rihanna
toyota	yvessaintlaurentbeautyuse	shakira
volvocars		taylorswift
		vindiesel

Table 1 shows all the names of the Facebook pages I collected which includes bmw Calillac, chevroelt, ford, gmc, honda, Hyundai, jeep, kia for car industry, armanibeauty Belifusa, cliniqueus, karitybeauty, lamer, laprairie, loccitaneusa, Maybelline, narscosmetics, origins, Sephora, yvessaintlaurentbeautyuse for cosmetic industry and I also collected data for several celebrities barackobama, brunomars, donaldtrump, dwaynejohnson, Jackie, justinbieber, katyperry. Ladygaga, leomessi, linkinpark, rihanna, Shakira, taylorswift, vindiesel.

Table 2 shows all the description and note of all variables. For example, the status_id is a bunch of digits consisting a unique number to represent one status, the status_message stands for the post text content in text character format, the status_type shows the type of post which includes photo, video and link and some other categories.

Table 2

Name	Description	Note
status_id	The status id	
status_message	The post text content	
link_name	The link name	
status_type	The post type	May contain photo/video/link
status_link	The url of link	
status_published	The timestamp of post published	
num_reactions	Num of reactions	Sum of likes, loves, wows, hahas, sads, angrys, specials
num_comments	Num of comments	Min = 0
num_shares	Num of shares	Min = 0
num_likes	Num of likes	Min = 0
num_loves	Num of loves	Min = 0
num_wows	Num of wows	Min = 0
num_hahas	Num of hahas	Min = 0
num_sads	Num of sads	Min = 0
num_angrys	Num of angrys	Min = 0
num_special	Num of special	Min = 0
text_characters	Text of characters	Min = 0
sentiment	Sentiment type	Positive or Negative
sentiment_score	Score of sentiment	Min = 0
lables	Lables of images	From Google vision API
fraction	The biggest fraction of one image	From Google vision API
tr	The red amount	From Google vision API
tg	The green amount	From Google vision API
tb	The blue amount	From Google vision API
faces	Num of faces	From Google vision API
landmarks	Num of landmarks	From Google vision API
logo	Num of logo	From Google vision API

The first part of data includes the posts content, time stamp, number of likes, number of shares, number of loves and so on. The second part of data adds the image analyzing results.

The names of the data included: status_idstatus_link, status_published, num_reactions, num_comments, num_shares, num_likes, num_loves, num_wows, num_hahas, num_sads,, num_angrys, num_special, text_characters, sentiment, sentiment_score, lables, fraction, tr, tg, tbfaces, landmarks, logo and the descriptions are recorded in Figure. To give an example to see how IBM Watson natural language processing works,

3.2 Data Exploration

3.2.1 Overview of Data

The raw data has 88991 rows times 21 columns of data.

Figure 4 Describing Dataset

```
   industry              name              status_id          status_message        link_name
 Length:88991       Length:88991       Length:88991       Length:88991       Length:88991
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



 status_type        status_link        status_published              num_reactions      num_comments
 Length:88991       Length:88991       Min.   :2008-01-30 17:47:19   Min.   :      0   Min.   :      0
 Class :character   Class :character   1st Qu.:2012-04-03 02:22:36   1st Qu.:    324   1st Qu.:     17
 Mode  :character   Mode  :character   Median :2014-03-31 08:09:15   Median :   1961   Median :     72
                                       Mean   :2014-03-01 02:35:28   Mean   :  48094   Mean   :   1616
                                       3rd Qu.:2016-02-22 11:03:36   3rd Qu.:  19177   3rd Qu.:    738
                                       Max.   :2018-03-08 18:04:41   Max.   :8247215   Max.   :1234260

   num_shares          num_likes          num_loves          num_wows             num_hahas
 Min.   :      0.0   Min.   :      0   Min.   :     0   Min.   :     0.00   Min.   :     0.00
 1st Qu.:      8.0   1st Qu.:    318   1st Qu.:     0   1st Qu.:     0.00   1st Qu.:     0.00
 Median :     77.0   Median :   1932   Median :     0   Median :     0.00   Median :     0.00
 Mean   :   1856.9   Mean   :  46499   Mean   :  1310   Mean   :    89.09   Mean   :    85.84
 3rd Qu.:    533.5   3rd Qu.:  18306   3rd Qu.:    20   3rd Qu.:     1.00   3rd Qu.:     0.00
 Max.   :1753584.0   Max.   :8044377   Max.   :504345   Max.   :116644.00   Max.   :180550.00

    num_sads            num_angrys         num_special         text_characters       sentiment
 Min.   :      0.0   Min.   :    0.0   Min.   :-7034.000   Min.   :-44620.0   Length:88991
 1st Qu.:      0.0   1st Qu.:    0.0   1st Qu.:    0.000   1st Qu.:    81.0   Class :character
 Median :      0.0   Median :    0.0   Median :    0.000   Median :   130.0   Mode  :character
 Mean   :     84.7   Mean   :   23.3   Mean   :    4.697   Mean   :   165.4
 3rd Qu.:      0.0   3rd Qu.:    0.0   3rd Qu.:    0.000   3rd Qu.:   203.0
 Max.   :1196402.0   Max.   :66284.0   Max.   :13741.000   Max.   : 17583.0
                                                           NA's   :16526
 sentiment_score
 Min.   :-35803.00
 1st Qu.:     0.00
 Median :     0.58
 Mean   :    -0.13
 3rd Qu.:     0.81
 Max.   :   144.00
 NA's   :16571
```

Figure 4 shows the summary of raw data which includes the data structure, the count of records, the levels and max, min values.

3.2.2 Data Clean

By reading the data summary, some abnormal and wrong data can be found. For example, num_special means the number of clicking special responses. Meanwhile, the text characters can be lower than 0. For the sentiment score, it should be between -1 and 1. The time stamp is difficult to use, so I used as.POSIXlt function to format the date and document the dates in day, month, year, etc. After cleaning data, the summary looks like:

```
> summary(all_withoutimage)
     industry          name             status_id         status_message       link_name
 car       :23330   Length:58523       Length:58523       Length:58523       Length:58523
 celebrity:19635   Class :character   Class :character   Class :character   Class :character
 cosmetic :15558   Mode  :character   Mode  :character   Mode  :character   Mode  :character



 status_type        status_link        status_published              num_reactions      num_comments
 Length:58523       Length:58523       Min.   :2008-02-08 14:57:11   Min.   :      0   Min.   :      0
 Class :character   Class :character   1st Qu.:2012-05-29 08:33:10   1st Qu.:    273   1st Qu.:     14
 Mode  :character   Mode  :character   Median :2014-05-19 13:38:04   Median :   1797   Median :     63
                                       Mean   :2014-04-12 13:05:58   Mean   :  39801   Mean   :   1572
                                       3rd Qu.:2016-03-18 16:28:42   3rd Qu.:  19144   3rd Qu.:    741
                                       Max.   :2018-03-08 18:04:41   Max.   :8247215   Max.   :1234260

   num_shares         num_likes          num_loves            num_wows            num_hahas
 Min.   :      0   Min.   :      0   Min.   :     0.0   Min.   :    0.00   Min.   :     0.00
 1st Qu.:      7   1st Qu.:    267   1st Qu.:     0.0   1st Qu.:    0.00   1st Qu.:     0.00
 Median :     68   Median :   1774   Median :     0.0   Median :    0.00   Median :     0.00
 Mean   :   1768   Mean   :  38663   Mean   :   913.6   Mean   :   58.79   Mean   :    81.45
 3rd Qu.:    500   3rd Qu.:  18306   3rd Qu.:    17.0   3rd Qu.:    1.00   3rd Qu.:     0.00
 Max.   :1753584   Max.   :8044377   Max.   :341089.0   Max.   :47889.00   Max.   :180550.00

    num_sads           num_angrys          num_special         text_characters       sentiment       sentiment_score
 Min.   :     0.0   Min.   :    0.00   Min.   :    0.000   Min.   :    8.0   negative: 9456   Min.   :-0.9773
 1st Qu.:     0.0   1st Qu.:    0.00   1st Qu.:    0.000   1st Qu.:   91.0   positive:49067   1st Qu.: 0.3011
 Median :     0.0   Median :    0.00   Median :    0.000   Median :  139.0                    Median : 0.6852
 Mean   :    56.4   Mean   :   23.18   Mean   :    4.202   Mean   :  179.5                    Mean   : 0.4900
 3rd Qu.:     0.0   3rd Qu.:    0.00   3rd Qu.:    0.000   3rd Qu.:  216.0                    3rd Qu.: 0.8371
 Max.   :597711.0   Max.   :66284.00   Max.   :13741.000   Max.   :17583.0                    Max.   : 0.9994


     mydate                          year            month             day               weekday
 Min.   :2008-02-08 14:57:11   2015   : 8097   10     : 5296   14     : 2009   Friday   :10036
 1st Qu.:2012-05-29 08:33:10   2014   : 7861   11     : 5289   6      : 1993   Monday   : 9361
 Median :2014-05-19 13:38:04   2016   : 7493   9      : 4998   23     : 1989   Saturday : 4957
 Mean   :2014-04-12 13:05:58   2017   : 7386   6      : 4905   11     : 1983   Sunday   : 4794
 3rd Qu.:2016-03-18 16:28:42   2013   : 7203   8      : 4896   20     : 1983   Thursday : 9697
 Max.   :2018-03-08 18:04:41   2012   : 7154   2      : 4859   18     : 1968   Tuesday  : 9775
                               (Other):13329   (Other):28280   (Other):46598   Wednesday: 9903
      hour             min              sec              yday
 11     : 6324   0      :11241   0      : 5175   44     :  225
 12     : 6126   30     : 2297   1      : 3676   145    :  211
 13     : 5040   1      : 1428   2      : 2072   59     :  209
 14     : 4944   2      : 1089   3      : 1653   334    :  206
 10     : 4722   15     : 1071   4      : 1578   22     :  205
 15     : 4661   5      : 1047   5      : 1388   306    :  204
 (Other):26706   (Other):40350   (Other):42981   (Other):57263
```

From the summary for cleaned data, the sentiment has two factors which are negative and positive, year, month, day, weekday, hour is all factorized and do not see any levels that should not appear.

3.2.3 Correlation Analysis

Figure 5 Correlation Analysis

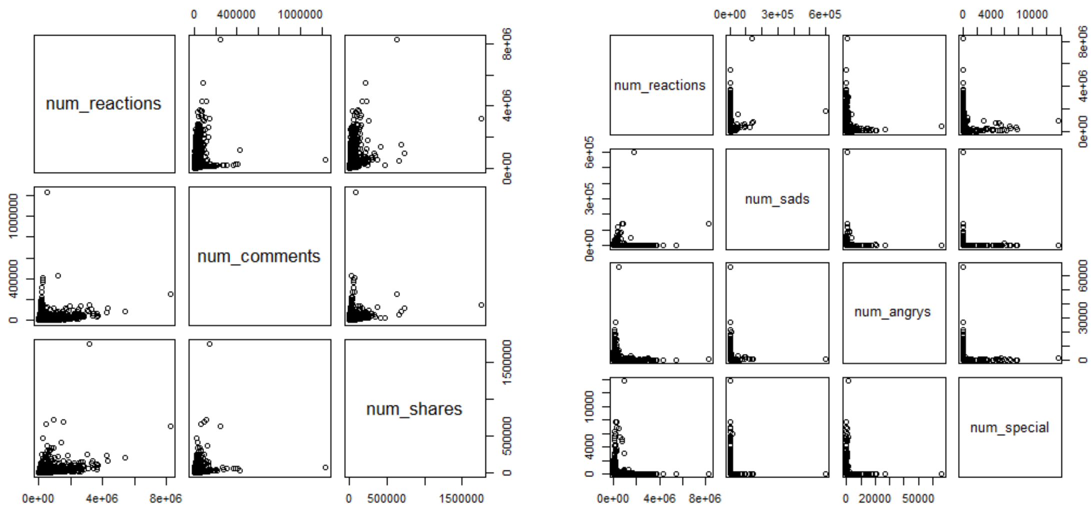


Figure 5 shows the correlation for number of reactions, comments, shares.

From the correlation analysis, we can tell that the correlations between number of reactions, number of comments, number of shares are not linear.

However, because the nature of the dataset, which includes the car industry, cosmetic industry and the data from several random celebrities, and this may cause the obvious two regression lines in the correlation graph. Therefore, to give a better understanding of the difference between different industries, I use the scatter diagram to show the relationships. Figure 6 shows the scatter diagram of the reaction and shares and comments. However, it can be found that there are still huge differences in the cosmetic dataset itself, to illustrate deeper, I am going to draw the smooth line for individual pages.

Figure 6 Relationship for all data

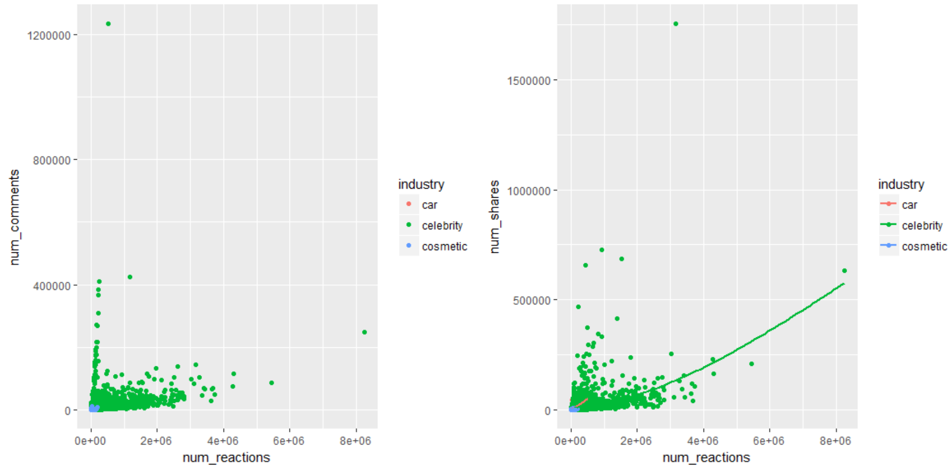


Figure 7, 8, 9 takes BMW and Lamer and Lady Gaga as three examples to see the relationships between reactions, shares, comments.

Figure 7 shows the BMW, from Figure 7, a positive relation can be found from the number of reactions and number of shares.

Figure 7 Relationship for BMW

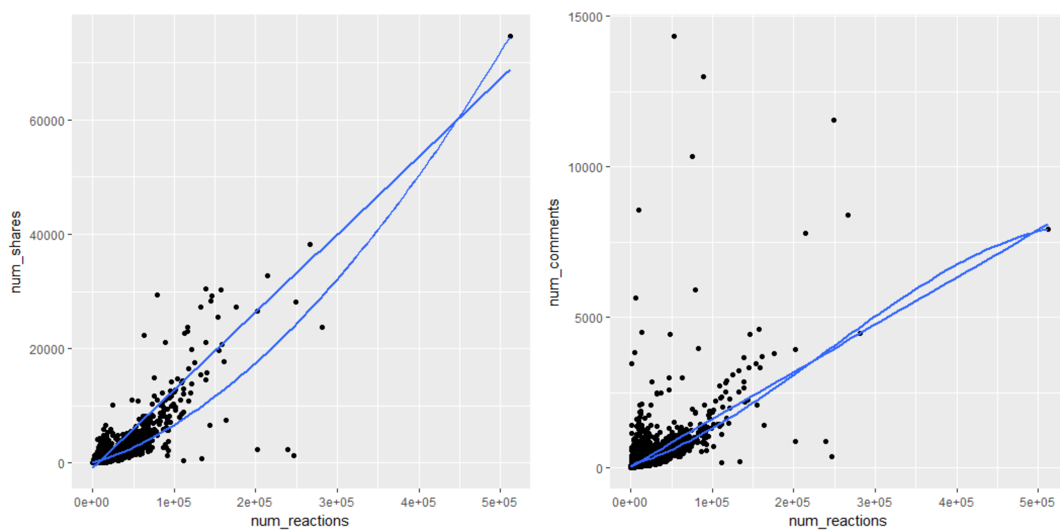


Figure 8 Relationship for Lamer

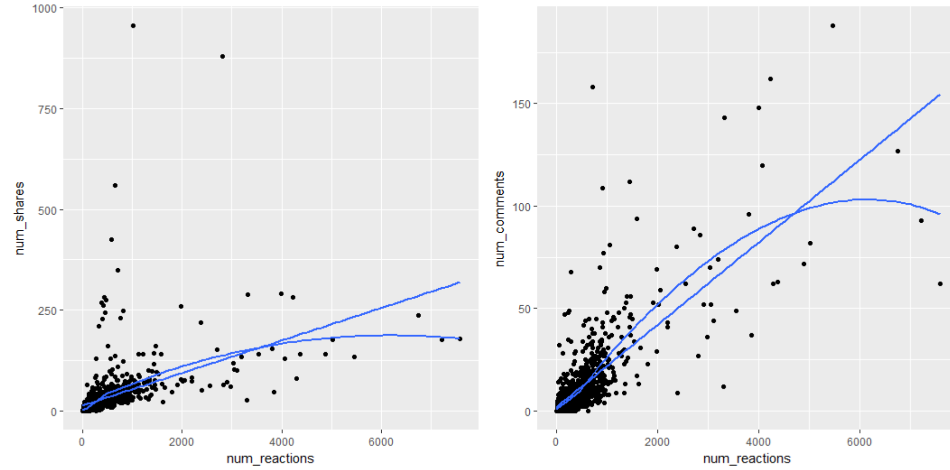
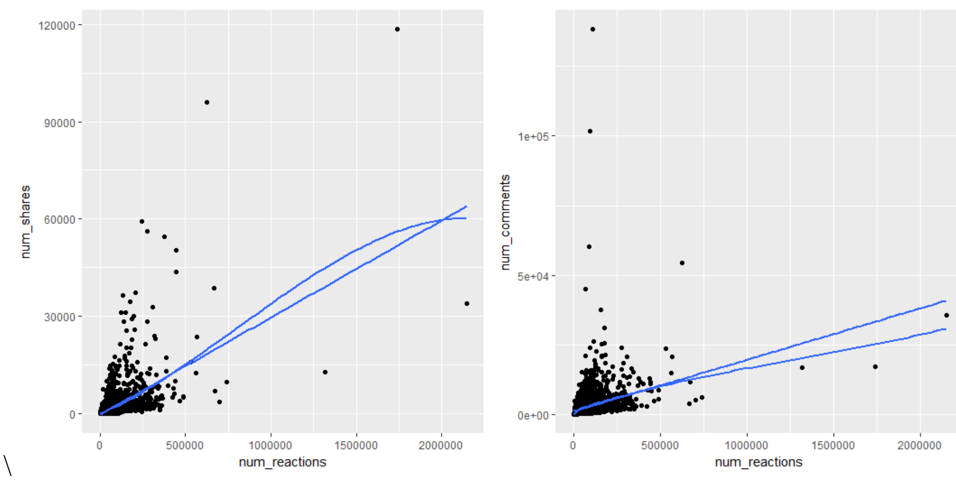


Figure 9 Relationship for Lady Gaga



\

Figure 8 shows the Lamer, from Figure 8, a positive relation can be found from the number of reactions and number of shares.

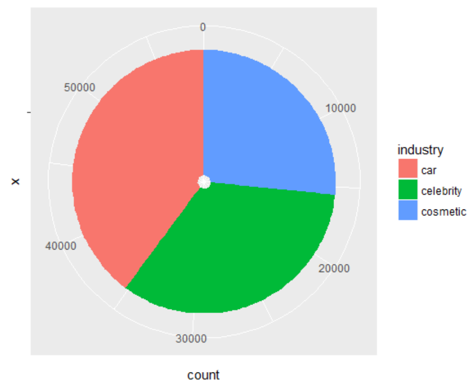
Figure 9 shows the Lady Gaga, from Figure 9, a positive relation can be found from the number of reactions and number of shares.

To conclude, the relations between reaction and comments and shares for different industries and different pages are greatly different, there is no such a universal conclusion that we can utilize to explain all pages, however, for one single page, there is a positive relation which means number of reactions and number shares are corelated.

3.2.4 Pattern Research

First of all, by counting the status_id, I can find that this dataset includes about one-fourth of cosmetic records and celebrity as well as car industry contributes about the same records.

Figure 10 The status different industry post



To evaluate what type of posts each industry prefers, the histogram is the most convincible diagram, and from Figure 11, it can be found that business post the most photos and some links as well as some videos. Meanwhile, the car and cosmetic industry believed photos more than the individuals. Surprisingly, celebrities love posting videos more than businesses.

Figure 11 The status different industry post

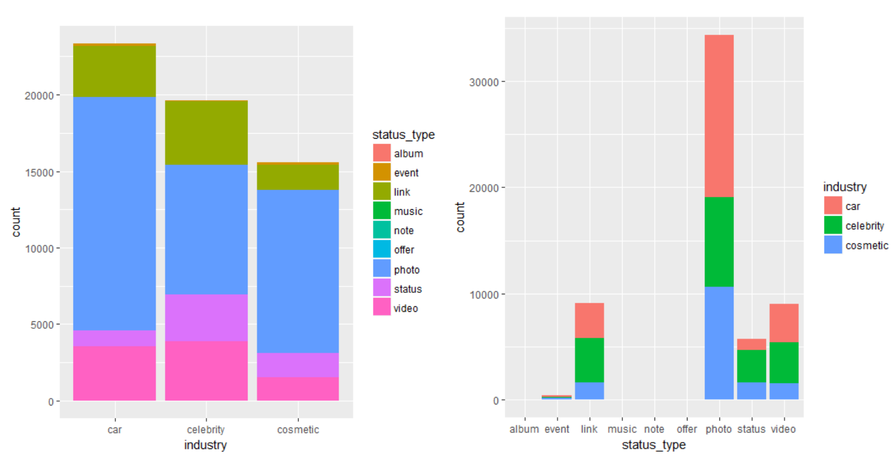
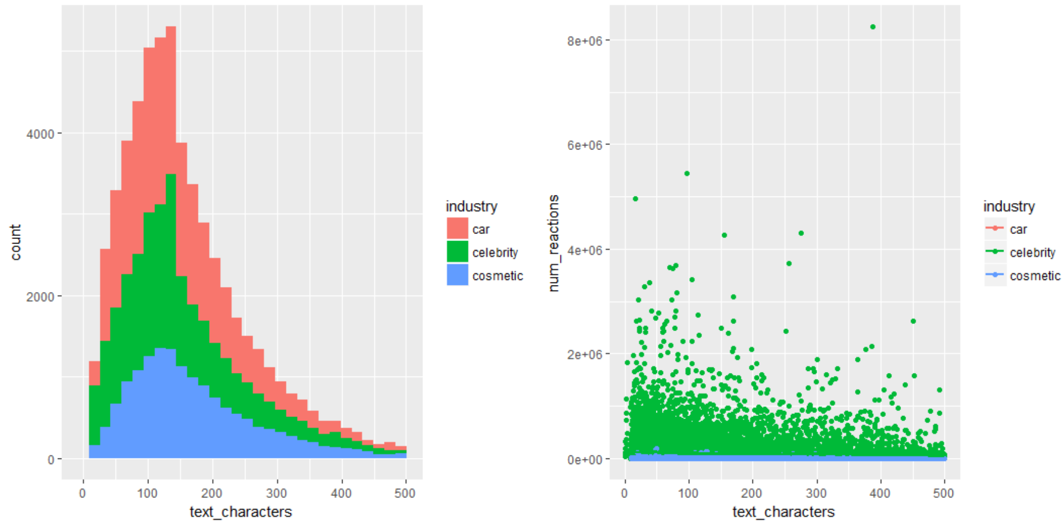


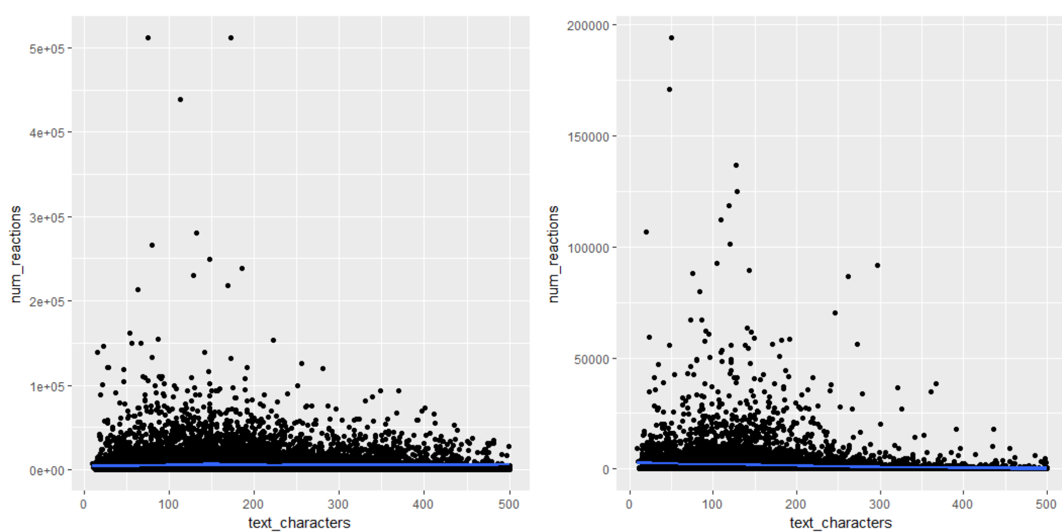
Figure 12 Histogram of text_characters



From the above histogram of text_characters, we can tell that most text_charaters are around 100 characters which is universal for all three industries and it seems that the number of reactions does not relate to text_characters, and the scatter diagram of text_charaters versus number of reactions does not seem to be very clear.

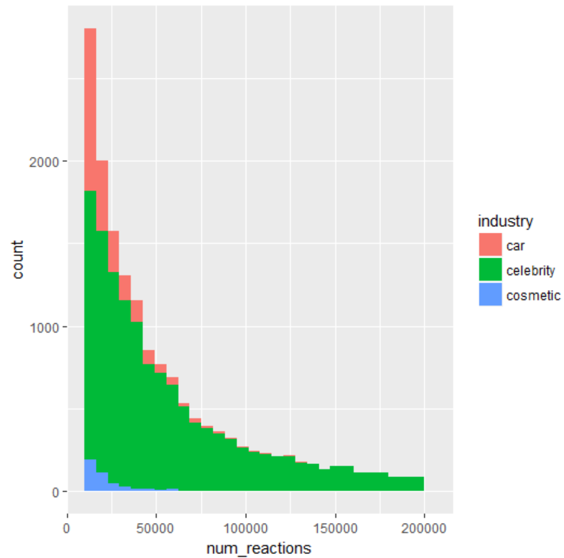
Therefore, I did two diagrams for the two industries. And from figure 13, which is respectively for car and cosmetic. A slightly negative relation could be observed. Which means that for most of time, the text_charaters close to 50-100 may attract more people or users to engage while more characters may keep people away.

Figure 13



To explore the nature of reactions, let's firstly dig into the frequencies of the reactions.

Figure 14 Histogram of number of reactions



From Figure 14, it can be found that celebrities have more reactions than the businesses do. And most posts have few reactions while fewer posts have reactions more than 100000.\

Figure 15 The relationships of number of reactions versus number of likes

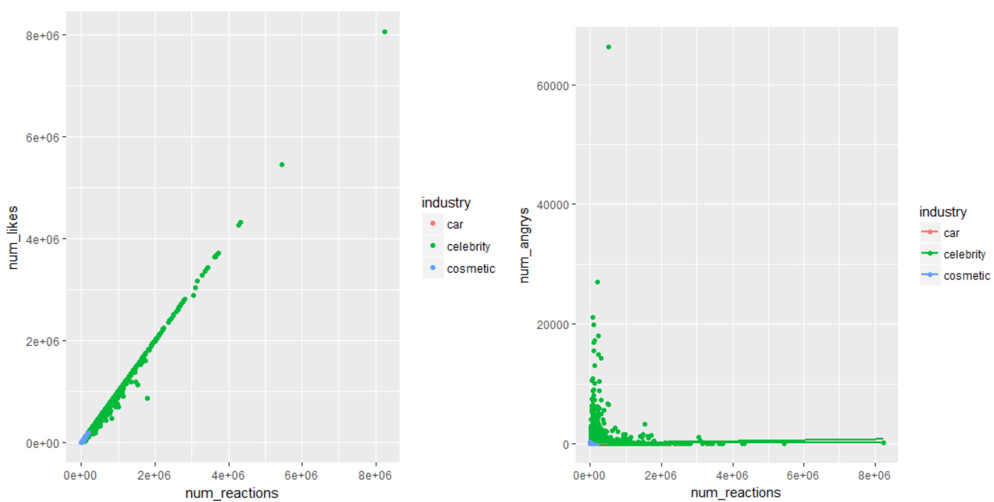
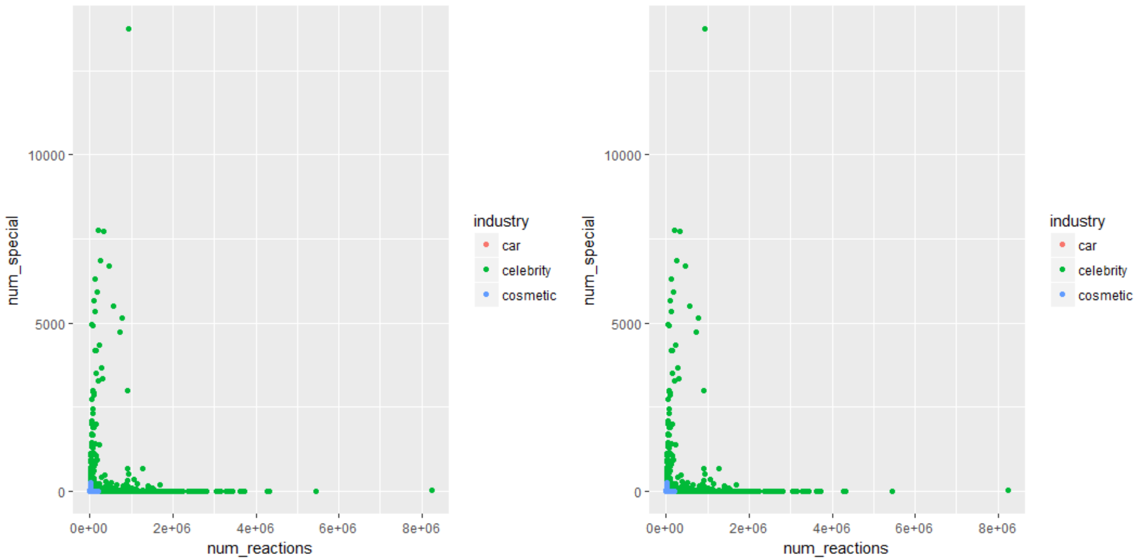


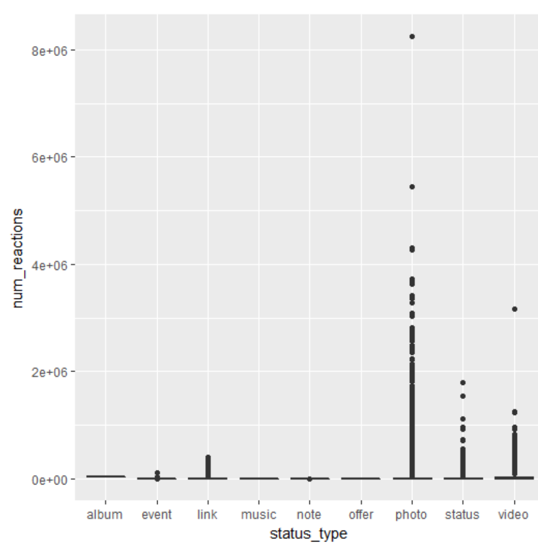
Figure 16



One of the interesting findings is that most reactios are consisted of likes, while some for number of angry, when the number of angry is high, the number of reaction is low, which ultimately means that peoples tend to have similar opinions on one post, seldom posts may seems to be controversial, therefore, for the rest of the project, I am going to use number of reactions as the target identifying the engagement of the users.

To identify the difference of number of reactions for different type of posts, the boxplot is a good option.

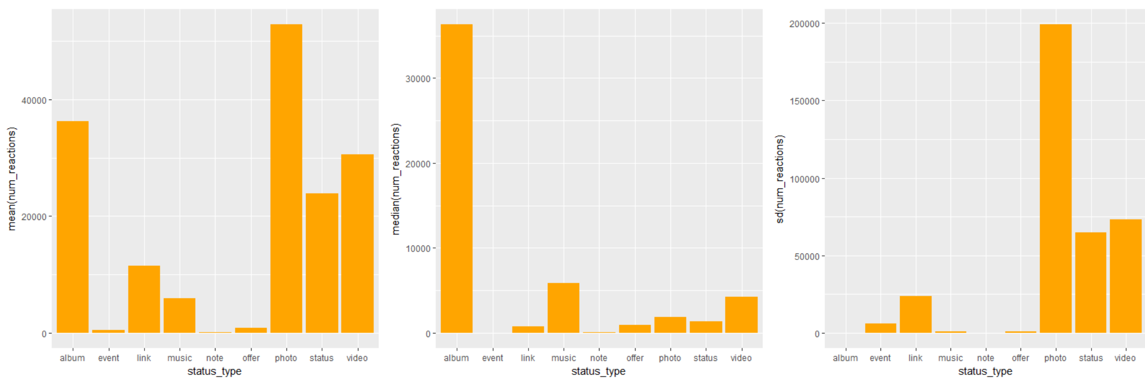
Figure 17



However, from Figure 17, because of the extensive external points, the specific pattern can not be discovered, therefore, Figure 18 which contains the means, medians and sds may explain more clearly.

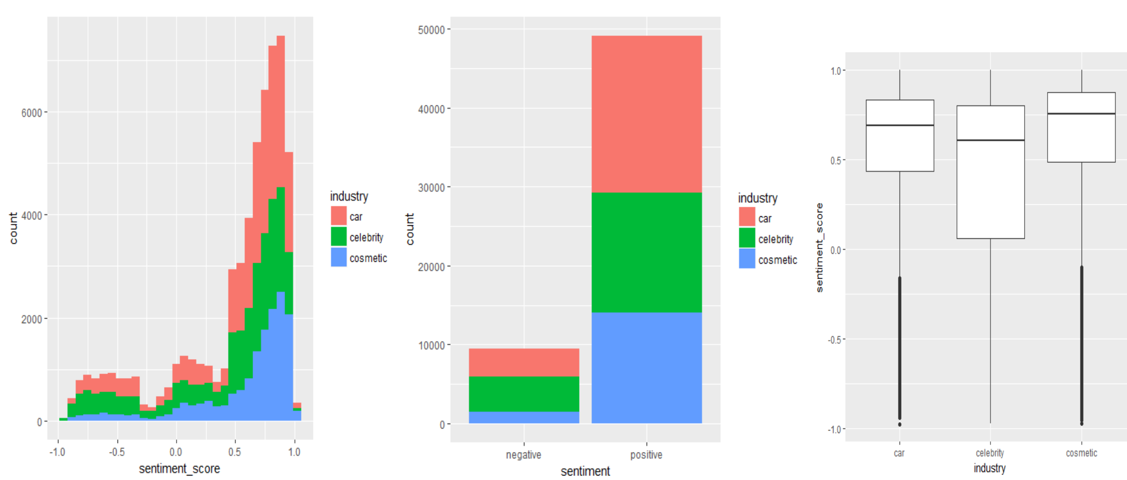
It can be found from the mean graph that photo has a very high mean value followed by album and video but for median, album and music seem to be very high, from Figure 11, we can find that the number of music and album are actually very low which makes the result inconvincible. Photo, status and video and links are four types with enough data we may analyze. The mean of photo is relatively high while the median of photo is lower than video which implies that the reaction numbers for photo is very diverse, may have huge difference from post to post while video is kind of with low standard deviation showing that most video contents seems to have similar and relatively high reactions.

Figure 18 Histogram of status_type versus mean, median, sd of number of reactions



To analyze the sentiment score, it can be found from Figure 19 that most posts have a relatively high sentiment score and a positive attitude, but it can be told that the celebrities are more likely to post negative posts.

Figure 19 Sentiment score



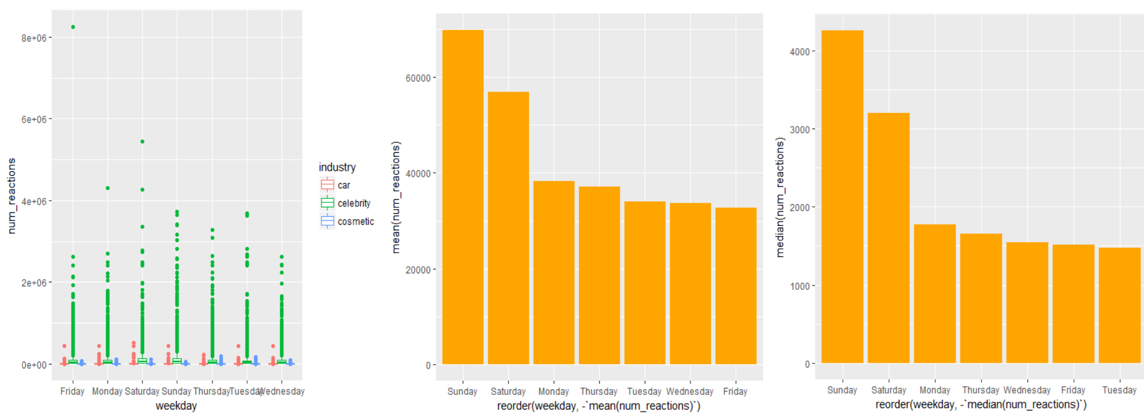
4 Analysis Approach and Solutions

4.1Time-Series

Time-Series is one of the most important parts of the analysis, before the exploration, here are some hypothesis: Is there one specific day or hour that people may engage more? Do the companies posts during the peak time?

From Figure 20, similarly, the box plot may not tell the entire story, because most numbers are close to zero, therefore, the mean and median diagrams are used and we can find that apparently, the number of reactions on Weekends are much more than weekdays which shows that people are more likely to use social media tools during weekends instead of weekdays.

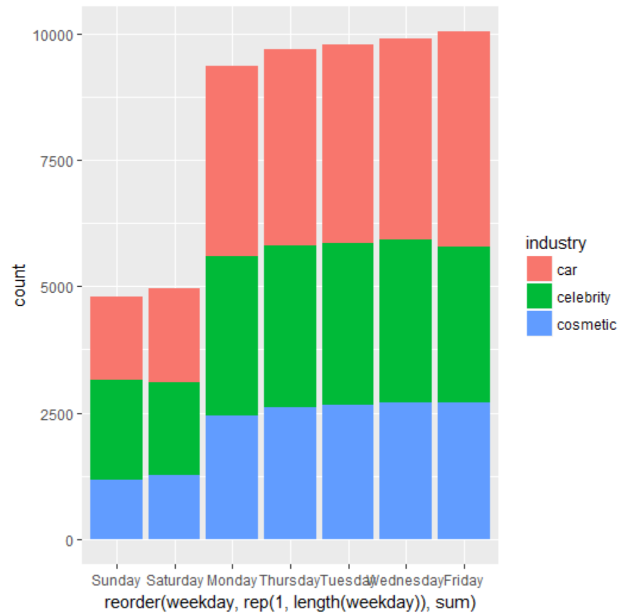
Figure 20 Weekday versus number of reactions



So, the question goes to does the pages runners, or operators posts more frequently on weekends comparing to on weekdays?

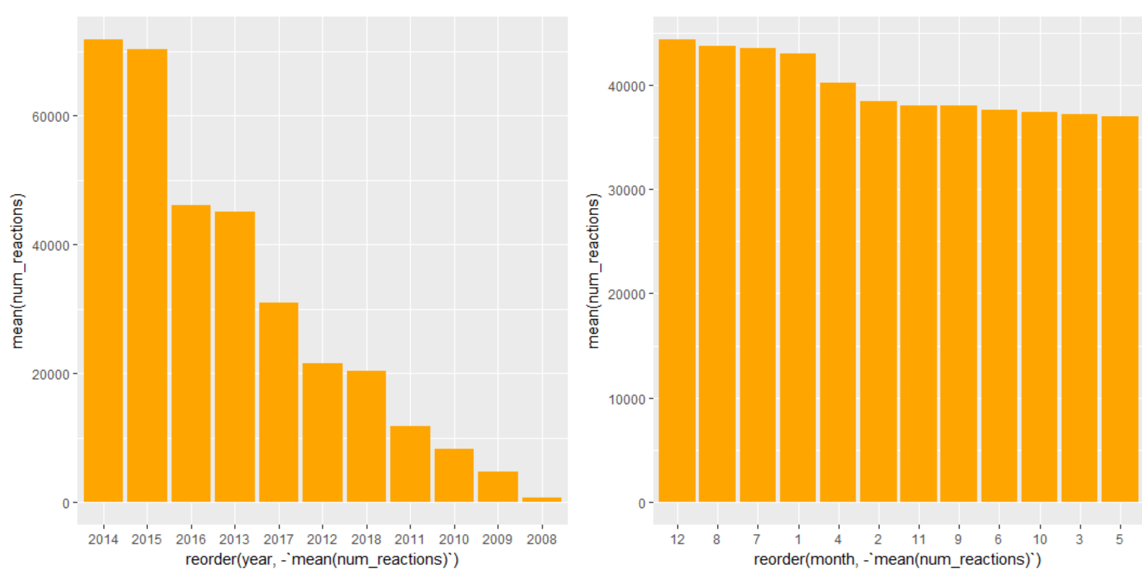
From Figure 21, the answer seems to be ‘no’. Exactly the reverse, the operators of Facebook pages are definitely employed by the business, so they are more likely to post on weekdays, not weekends. The recommendation here is going to be post more on weekends, not weekdays, if the business is looking for more engagement from users.

Figure 21



Similarly, Figure 22 shows the year & month versus number of reactions, from Figure 22, it can be discovered that the business sees a higher engagement from 2014 to 2015, but see a great drop since 2016, this may due to couple of reasons including the emergence of other social media tools like Instagram, Tumbler or Twitter. People may transfer to these platforms instead of Facebook pages, and from the monthly data, there are no explicit differences from one month to another.

Figure 22



One more interesting finding is for hours, from figure 23, we can tell that most reactions happens on midnight or late night, but from figure 24, it is obvious that business page runners post more frequently from 10 to 16 and like weekday analysis, the recommendations should be post more frequently from 22 to 1 because the users are more active and are more likely to engage with what business posts.

Figure 23 hour versus reaction

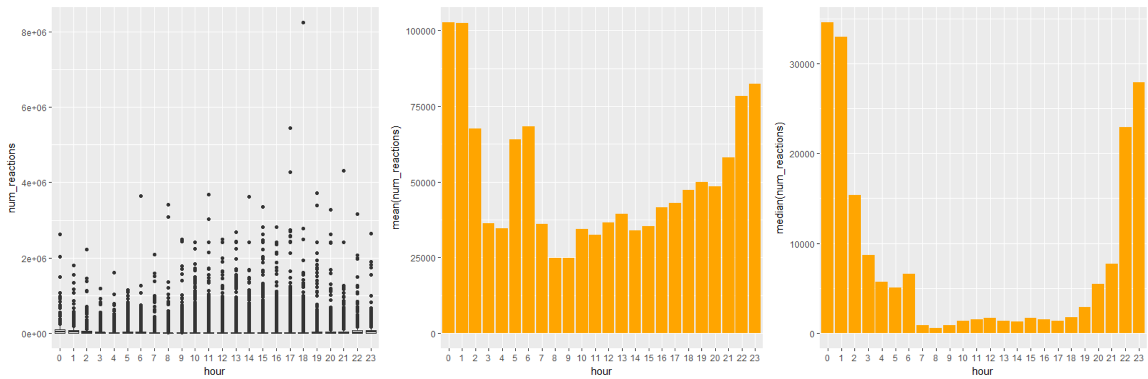


Figure 24

4.2 Linear Regression

To better discuss the effects from different aspects of data on the number of reactions, I run the linear regression for the several aspects.

Linear regression is a statistical tool used to model the relation between a set of characteristics and targets. It estimates the mean value of the response variable, given levels of the predictor variables. The regression approach complements the least squares by identifying how differently targets respond to a change in one unit of characteristic, rather than estimating the constant regression coefficient representing the change in the response variable produced by a one-unit change in the predictor variable associated with that coefficient. It estimates the implicit price for each characteristic across the distribution.

From Figure 25 to Figure 27, the regression model for different segments, respectively from text, type, to all. It can be found that the prediction model is not working as well as expected.

Figure 25 text-content segment

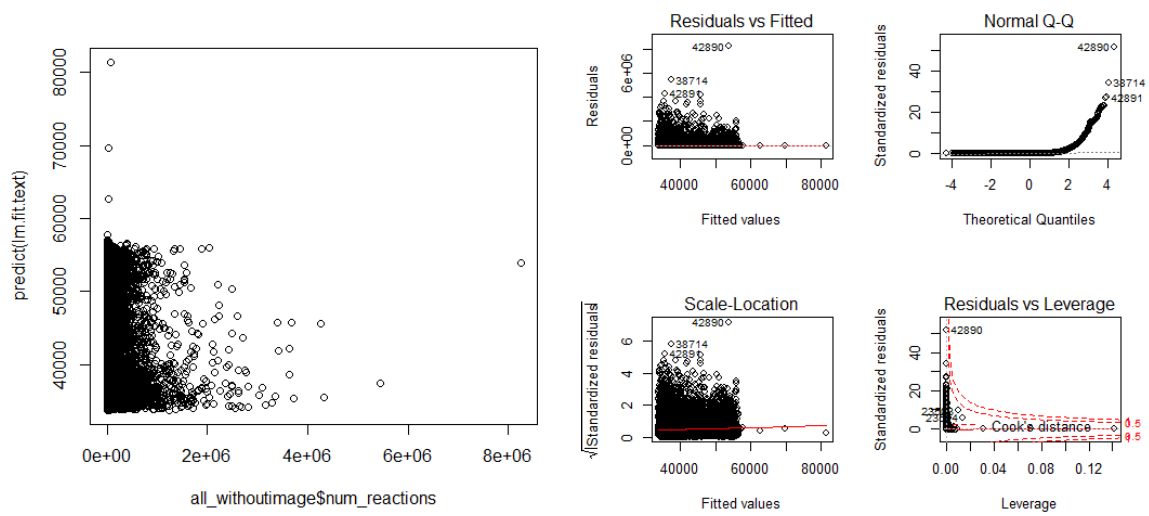


Figure 26 type and industry

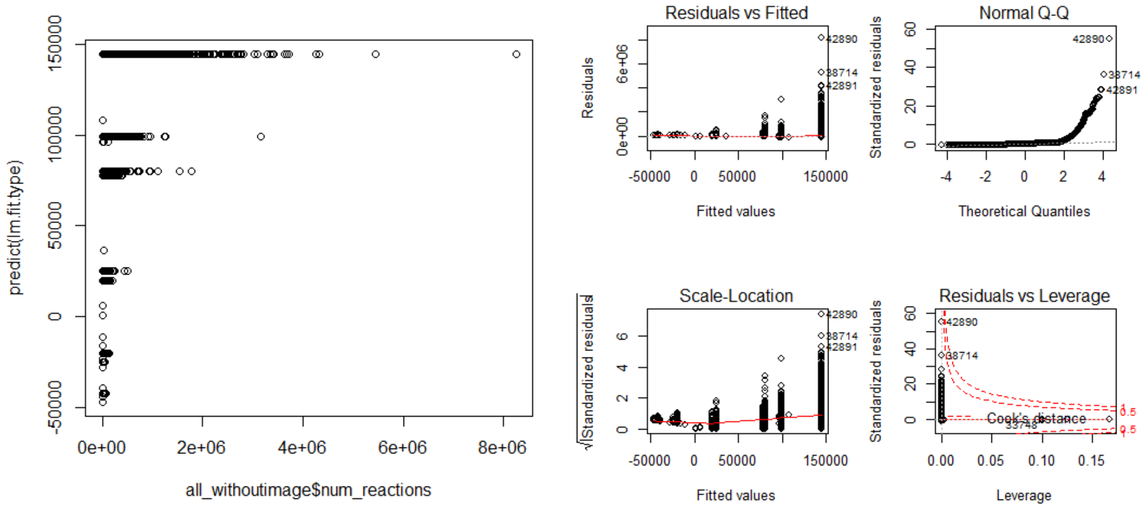
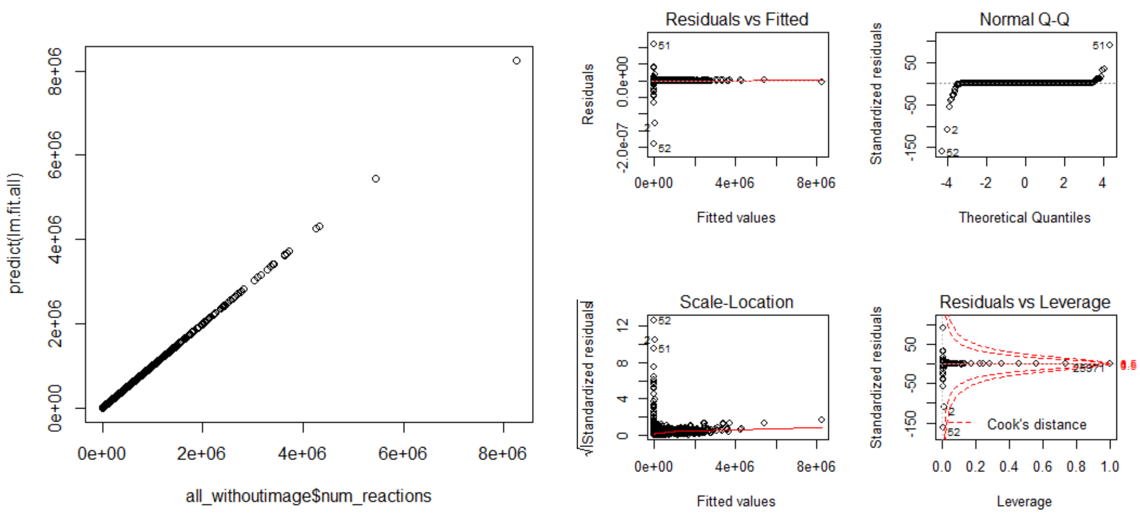
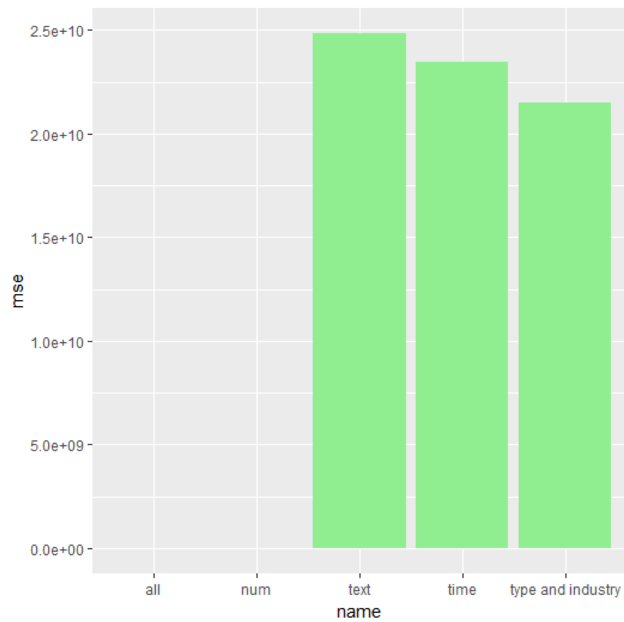


Figure 27 all features



To summarize, from figure 28, we can find that the most relevant feature segment is the number of like, number of angry, which is convincible and followed by type and industry, which means type and industry has a stronger impact on the number of reactions. Between text and time, time seems to be more relevant to the number of reaction

Figure 28



4.3Analyzing Image information & Text mining

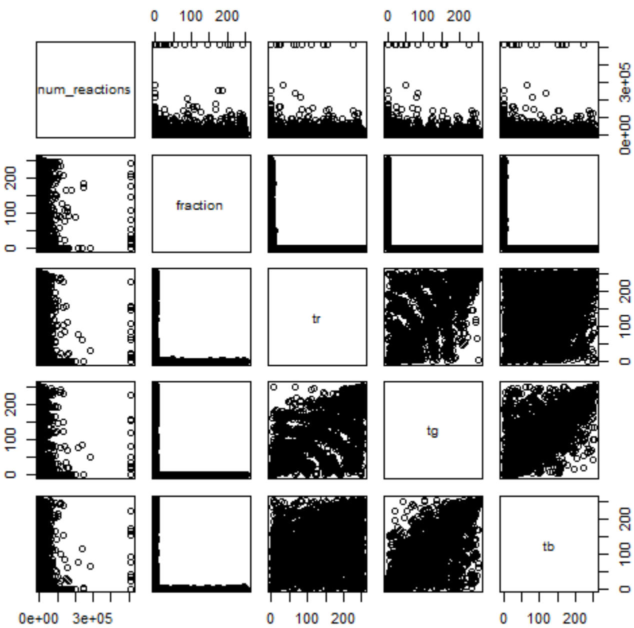
The Google Vision API helped me to extract the key words from images uploaded. For example, for such an image, the Google Vision API will return:

Figure 29

lables	fraction	tr	tg	tb	faces	landmarks	logo
['car', 'auto show', 'motor vehicle', 'personal luxury car', 'vehicle', 'sports car', 'automotive design', 'performance car', 'luxury vehicle', 'concept car']	0.077908	25	54	81	0	0	0

The labels are ['car', 'auto show', 'motor vehicle', 'personal luxury car', 'vehicle', 'sports car', 'automotive design', 'performance car', 'luxury vehicle', 'concept car']. By text mining the labels, I may have the opportunity to find what images people may want to take a look. After cleaning all the data, deleting those with 0 tg, tr and tb, I finally collected 6980 rows times 29 columns of data for image information.

Figure 30



From the given dataset, there is no strong relationships between the color and the number of reactions. Some questions may be asked, like: Are there any images that may make people angry/laugh? What images may attract people?

This part is going to split into two parts, the first is car industry.

Figure 31

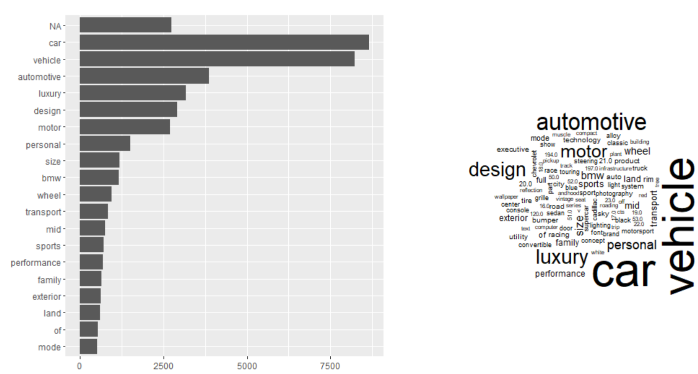
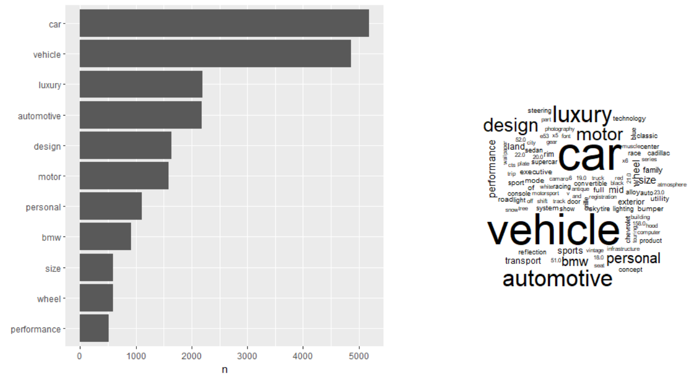


Figure 32



By comparing figure 31 and figure 32, which are respectively the numbers for all data and the figure selecting data whose reactions are more than average, we can tell that what kind of images may attract more people. From Figure 31 and Figure 32, some patterns may excite the businesses, for example, the term ‘luxury’ seems to attract more people.

Similarly, Figure 33 and Figure 34 compares cosmetic industry and from 33 and 34 ,we can definitely tell that the images with the tag ‘health’ are more likely to attract people.

Figure 33

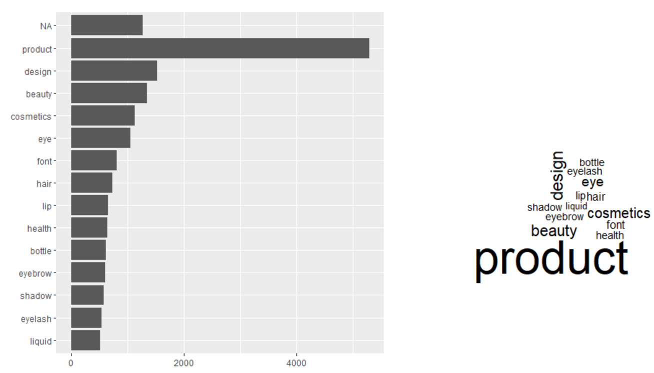
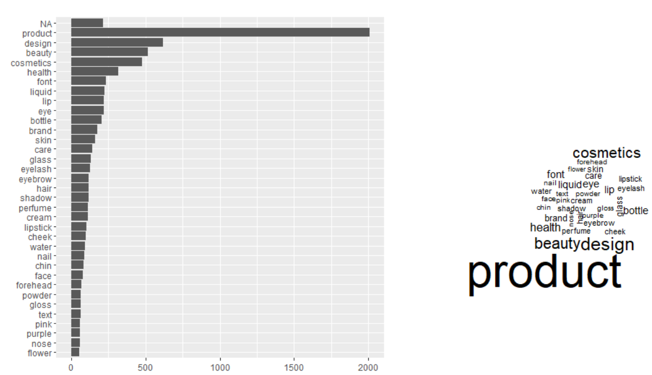


Figure 34



5 Conclusion

“The aim is to make users talk.” is what this project is trying to make it reality.

To conclude the findings, the relations between reaction and comments and shares for different industries and different pages are greatly different, there is no such a universal conclusion that we can utilize to explain all pages, however, for one single page, there is a positive relation which means number of reactions and number shares are corelated.

Here are some findings with recommendations: The text_charactes better be controlled from 50 to 150 to make more people engaged. Business should consider posting more frequently on weekends and late nights to attract more people to engage.

For car industry, it can be analyzed that the more ‘luxury’ tag may trigger more engagement while for cosmetic industry, people are more eager to hear or see something related to the tag ‘health’.

Though the project may not include all the data needed to give a broad view of Facebook page, some more steps including getting more images, building single model for individual page, analyzing the videos posted may help to give a better view.

References

Anderson, B., Fagan, P., Woodnutt, T., & Chamorro-Premuzic, T. (2012). Facebook psychology: Popular questions answered by research. *Psychology of Popular Media Culture,1*(1), 23-37. doi:10.1037/a0026452

Hale, T. M., Pathipati, A. S., Zan, S., & Jethwani, K. (2014). Representation of Health Conditions on Facebook: Content Analysis and Evaluation of User Engagement. *Journal of Medical Internet Research,16*(8). doi:10.2196/jmir.3275

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research,65*(10), 1480-1486. doi:10.1016/j.jbusres.2011.10.014

Lee, K., & Lim, H. (2017). Facebook Me Right: Needs-Based Segmentation of Facebook Brand Page Users. *Fashion, Industry and Education,15*(1), 12-28. doi:10.7741/fie.2017.15.1.012

Wu, Y. J., Chang, W., & Yuan, C. (2015). Do Facebook profile pictures reflect user's personality? *Computers in Human Behavior,51*, 880-889. doi:10.1016/j.chb.2014.11.014