Assignment3-Pratik-COVID_Vaccine_Sentiment.Rmd

Pratik Chaudhari

4/14/2021

Data Source:

https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

Background:

COVID-19 is an infectious disease caused by a newly discovered strain of coronavirus, a type of virus known to cause respiratory infections in humans. This new strain was unknown before December 2019. Ever since the Covid-19 pandemic there has been quite a buzz in social media platforms and news sites regarding the need for COVID-19 Vaccine. Hence, the data consist of recent tweets about COVID - 19 vaccines used in entire world on large scale, as following:

- Pfizer/BioNTech
- Sinopharm
- Sinovac
- Moderna
- Oxford/AstraZeneca
- Covaxin
- Sputnik V.

We will be focusing on the sentiments or the emotions of the people post taking the shots for COVID vaccine.

Variables:

- **id:** User ID
- **user_name:** username of the user
- **user_location:** Location of the User
- **user_description:** Description of the User
- **user_created:** Date of user account creation
- **user_followers:** Number of followers of the User
- **user_friends:** Number of friends of the User

- **user_favorites:** Number of Users marked favorite by the account holder
- **user_verified:** Is the User a verified User or not
- **date:** Date when the tweet was tweeted
- **text:** Text of the tweet
- **hastags:** Hashtags used in the tweets
- **source:** Source device from where the tweet was tweeted
- **retweets:** Count of retweets for the shared tweet

Loading the necessary Libraries:

```
library(qdap)

## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current
timezone 'T':
## please set environment variable 'TZ'

library(tidytext)
library(tidyr)
library(ggplot2)
library(dplyr)
library(stringr)
library(dplyr)
library(cowplot)
library(lubridate)
library(vader)
library(corpus)
library(SnowballC)
library(twitteR)
library(tm)
library(wordcloud)
library(syuzhet)
library(knitr)
library(naniar)
library(GGally)
```

Reading the data from .csv to r:

```
# Reading the data and Converting blank spaces to NA

testdata <- read.csv("vaccination.csv", header=T, na.strings=c("","NA"))

kable(testdata[1:5,], caption = "Dataframe")
```

Dataframe

id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	favorites	is_retweet

		on		ted	ers	nds	tes	ed				ce	ets	tes	eet
1.340539e+18	Rachel Roh	La Crescenta-Montrose, CA	Aggregator of Asian American news; scanning diverse sources 24/7/365. RT's, Follows and 'Likes' will fuel me ðŸ‘©â€ðŸ»	2009-04-08 17:52:46	405	1692	3247	False	2020-12-20 06:06:44	Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF	['PfizerBioNTech']	Twitter for Android	0	0	False
1.338159e+18	Albert Fong	San Francisco, CA	Marketing dude, tech geek, heavy metal & '80s music junkie. Fascinated by meteorology and all things in the cloud. Opinions are	2009-09-21 15:27:30	834	666	178	False	2020-12-13 16:27:13	While the world has been on the wrong side of history this year, hopefully, the biggest vaccination effort	NA	Twitter Web App	1	1	False

			my own.							we've evâ€¦ https://t.co/dlCHrZjkhm					
1.337858e+18	eliðŸ‡±ðŸ‡¹ðŸ‡ªðŸ‡ºðŸ‘Œ	Your Bed	heil, hydraðŸ–â˜	2020-06-25 23:30:28	10	88	155	False	2020-12-12 0:33:45	#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years â€¦ https://t.co/ieYlCKBr8P	['coronavirus', 'SputnikV', 'AstraZeneca', 'PfizerBioNTech', 'Moderna', 'Covid_19']	Twitter for Android	0	0	False
1.337856e+18	Charles Adler	Vancouver, BC-Ca	Hosting "CharlesAdlerTonight" Global News	2008-09-10 11:2	49165	3933	21853	True	2020-12-12	Facts are immutable, Senator, even when	NA	Twitter We	446	2129	False

		na da	Radio Network. Weeknights 7 Pacific-10 Eastern - Email comments/ideas to charles@charlesadlertonight.ca	8:53					20:23:59	you’re not ethically sturdy enough to acknowledge them. (1) You were born iâ€¦ https://t.co/jqgV18kch4		b App				
1.337854e+18	Citizen News Channel	NA	Citizen News Channel bringing you an alternative news source from citizen journalists that haven’t sold out. Real news & real views	2020-04-23 17:58:42	152	580	1473	False	2020-12-12 20:17:19	Explain to me again why we need a vaccine @BorisJohnson @MattHancock #whereareallthesickpeople #PfizerBioN	[‘whereareallthesickpeople’, ‘PfizerBioNTech’]	Twitter for iPhone	0	0	False	

Techâ
€¦
https:
//t.co
/KxbS
RoBE
Hq

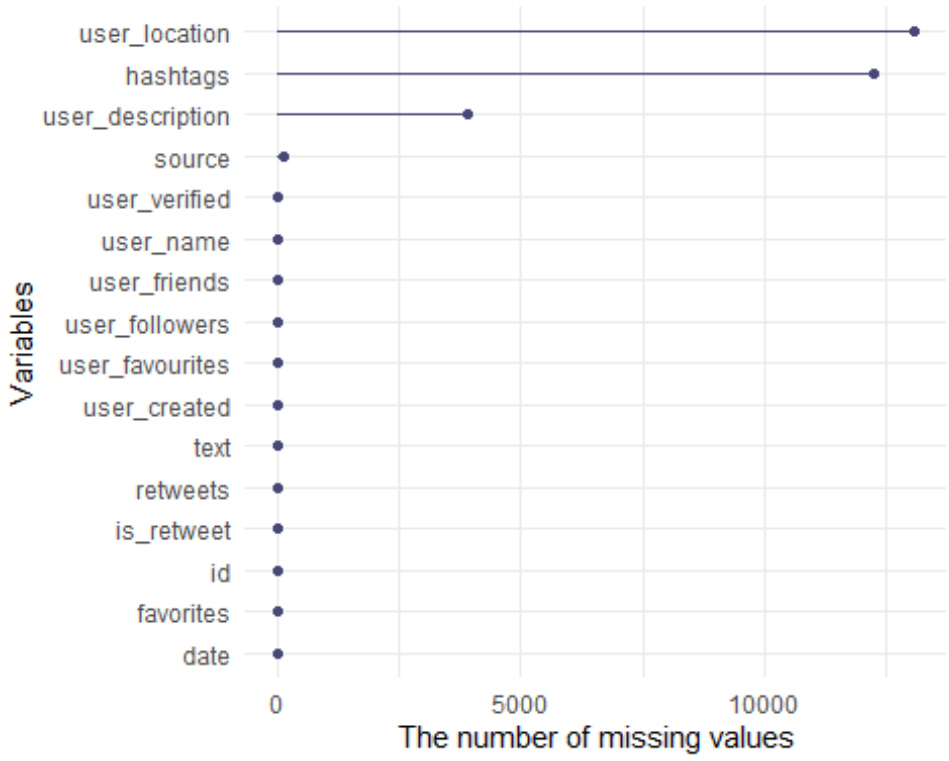
Data Cleaning:

- Here we can see the variables, and can have some idea about their type and what they contain. First, we will try to see if there are any missing values in this dataset. For that a 'naniar' library is loaded.
- Let's see the number of missing values for each variables and then plot the graphs for better visualization. In the first graph the number of missing values are plotted, in the 2nd one the percentage of the same are visualized.

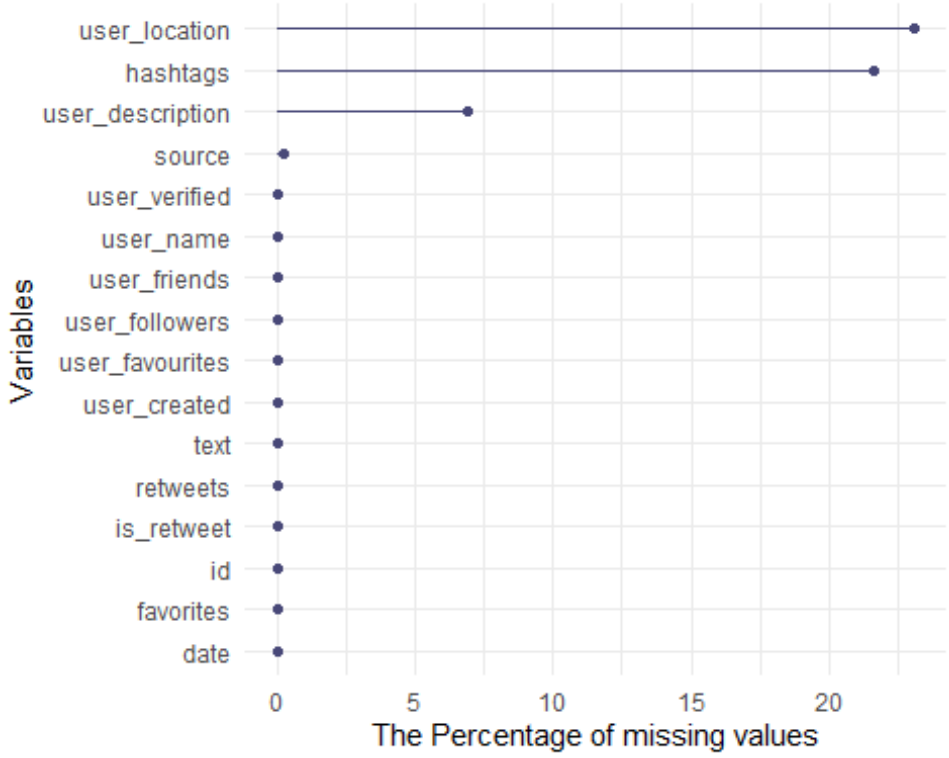
```
colSums(is.na(testdata))
```

```
##               id        user_name    user_location user_description
##                0                0            13086             3885
##     user_created   user_followers     user_friends  user_favourites
##                0                0                0                0
##    user_verified             date             text         hashtags
##                0                0                0            12238
##           source         retweets        favorites       is_retweet
##              125                0                0                0
```

```
gg_miss_var(testdata) + labs(y = "The number of missing values")
```




```
gg_miss_var(testdata, show_pct = TRUE) + labs(y = "The Percentage of missing values")
```



Dropping the **NA** values for getting better and more accurate results.

```
testdata = na.omit(testdata)

kable(testdata[1:3,], caption = "Dataframe after dropping 'NA")
```

Dataframe after dropping 'NA

	id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	favorites	is_retweet
1	1.340539e+18	Rachel Roh	La Crescenta-Montrose, CA	Aggregator of Asian American news; scanning diverse sources 24/7/365. RT's, Follows and 'Likes' will fuel me ðŸ‘	2009-04-08 17:52:46	405	1692	3247	False	2020-12-20 06:06:44	Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF	['PfizerBioNTech']	Twitter for Android	0	0	False

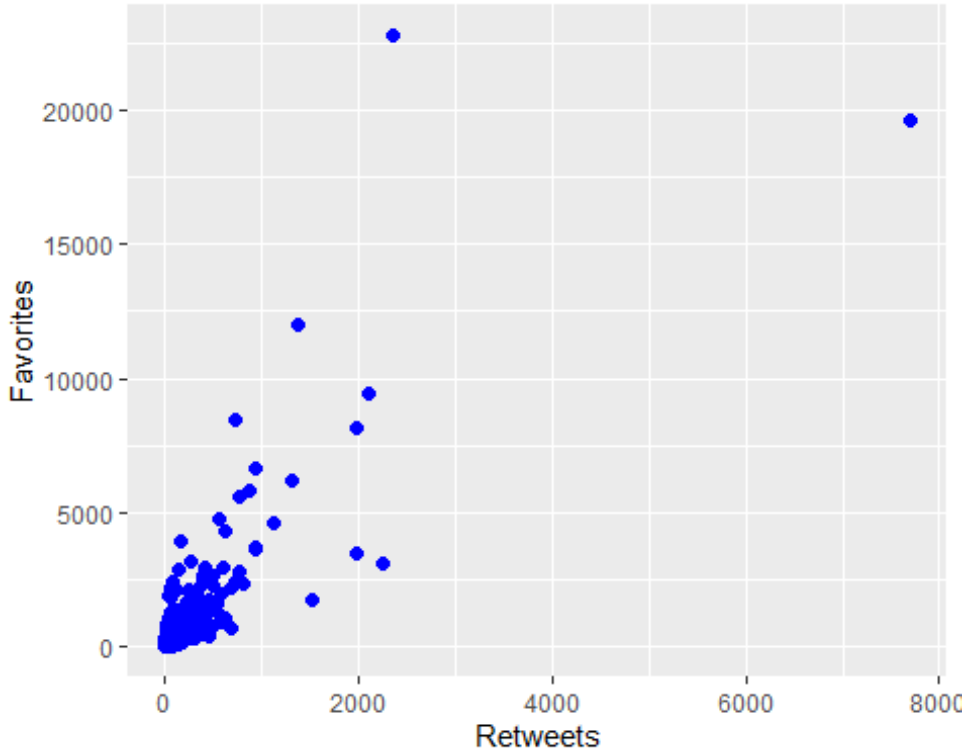
				©â€•ðŸ’»												
3	1.337858e+18	eliðŸ‡±ðŸ‡¹ðŸ‡ªðŸ‡ºðŸŒ	Your Bed	heil, hydra ðŸ–â˜º	2020-06-25 23:30:28	10	88	155	False	2020-12-12 20:33:45	#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years â€¦ https://t.co/ieYlCKBr8P	['coronavirus', 'SputnikV', 'AstraZeneca', 'PfizerBioNTech', 'Moderna', 'Covid_19']	Twitter for Android	0	0	False
7	1.337851e+18	Gunther Fehlinger	Austria, Ukraine and Kosovo	End North Stream 2 now - the pipelin	2013-06-10 17:49:22	2731	5001	6934	False	2020-12-12 20:0	it is a bit sad to claim the fame for success of #vaccinatio	['vaccination']	Twitter Web App	0	4	False

	e of cor rup tion , fun din g Rus sias war agai nst Ukr ain e,G eor gia, Syri a and poli tica l inte rve ntio n in USA and EU mu st be sto ppe d	6: 0 0	n on patrio tic comp etitio n betwe en USA, Canad a, UK andâ€¦ https: //t.co /IfMr AyGy TP

Data Visualization:

Let's try to see whether there's any correlation between the number of retweets and favorites. Because, in reality, we would assume to have a strong correlation between them.

```
ggplot(testdata, aes(x=retweets, y=favorites)) + geom_point(size=2,
color='Blue') +
  labs(x="Retweets", y="Favorites")
```

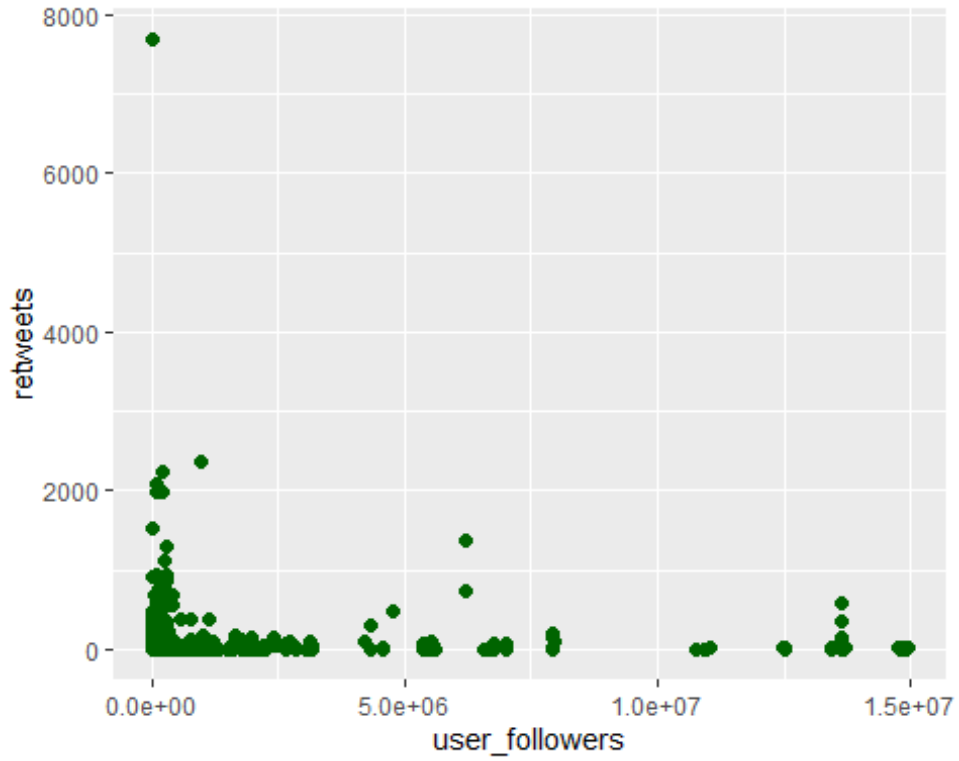


From the above graph we can observe that there is somewhat strong relation between the retweets and favorites. Lets confrim it with the cor() function.

```
cor(testdata$retweets, testdata$favorites)

## [1] 0.8351264
```

Now, Let's check the correlation between the users followers and retweets

```
ggplot(testdata, aes(x=user_followers, y=retweets)) + geom_point(size=2,
color='Dark green')
```




```
cor(testdata$user_followers, testdata$retweets)
```

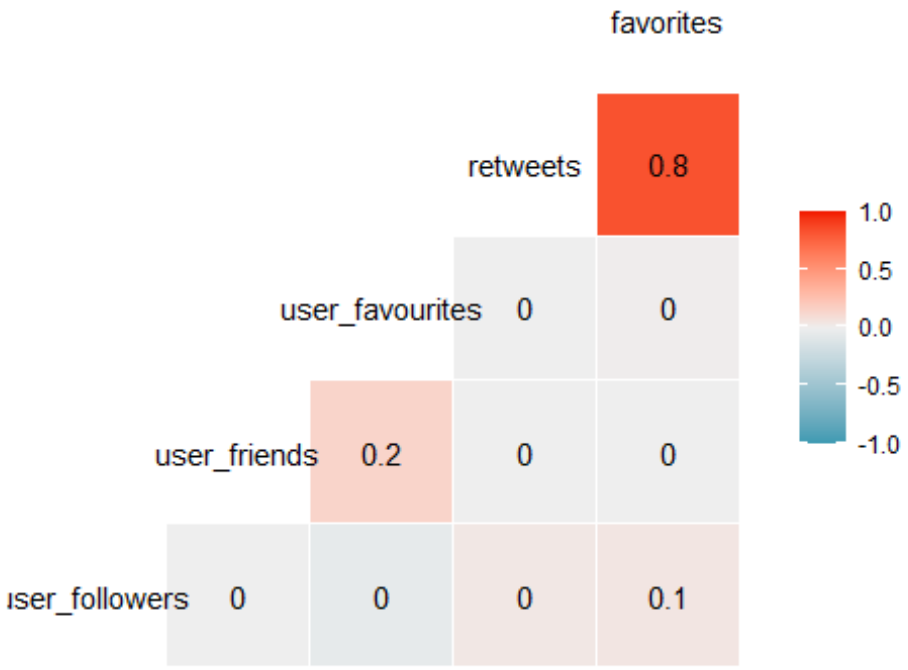
```
## [1] 0.04089292
```

From this, we can conclude that, having more followers and having more retweets have almost no correlation. So there are chances that, the tweets were Retweeted due to some other reasons. Like quality or for having same/relevant hashtags etc.

Now, Let's check the correlation matrix for all the numerical variables, excluding the id as it has no relevance.

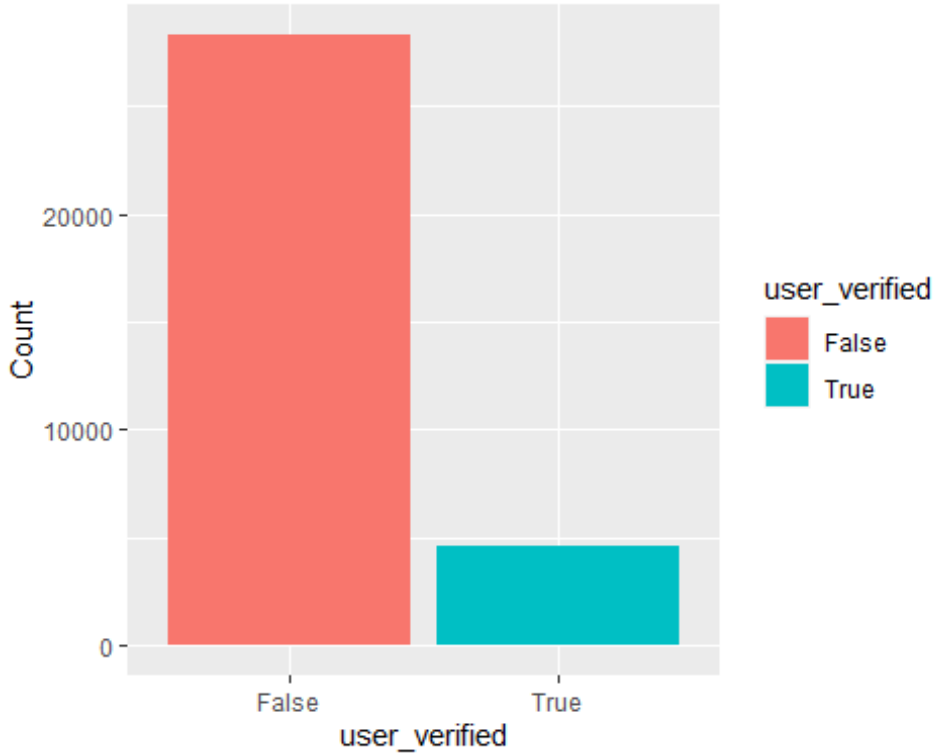
```
ggcorr(testdata[,-1], palette = "RdGy", label = TRUE)
```

```
## Warning in ggcorr(testdata[, -1], palette = "RdGy", label = TRUE): data in
## column(s) 'user_name', 'user_location', 'user_description',
'user_created',
## 'user_verified', 'date', 'text', 'hashtags', 'source', 'is_retweet' are
not
## numeric and were ignored
```



Now, Let's plot the count of verified and un-verified users

```
ggplot(testdata, aes(x=user_verified)) + geom_bar(aes(fill=user_verified)) +
  labs(y="Count")
```



As expected. Most users aren't verified.

Now, Let's check from which platform (source) the tweets were made. Let's first check how many total sources are there.

```
unique(testdata$source)
```

```
##   [1] "Twitter for Android"
##   [2] "Twitter Web App"
##   [3] "Twitter for iPhone"
##   [4] "TweetDeck"
##   [5] "Buffer"
##   [6] "Twitter for iPad"
##   [7] "LinkedIn"
##   [8] "Twitter for Mac"
##   [9] "24liveblog"
##  [10] "SocialFlow"
##  [11] "Instagram"
##  [12] "Socialbakers"
##  [13] "Echobox"
##  [14] "Microsoft Power Platform"
##  [15] "Hootsuite Inc."
##  [16] "Sendible"
##  [17] "Twitter Media Studio"
##  [18] "Nonli"
##  [19] "EastMojo"
##  [20] "Twitter Media Studio - LiveCut"
##  [21] "Tweetbot for Mac"
```

```
##  [22] "GT_Backend"
##  [23] "Sprout Social"
##  [24] "IFTTT"
##  [25] "TweetCaster for Android"
##  [26] "UberSocial for Android"
##  [27] "Blog2Social APP"
##  [28] "WordPress.com"
##  [29] "Paper.li"
##  [30] "Tweetbot for iÎŸS"
##  [31] "Sprinklr Publishing"
##  [32] "Twidere for Android"
##  [33] "Threader_client"
##  [34] "IndiaPost"
##  [35] "China Xinhua News"
##  [36] "Sqwarkr"
##  [37] "StockTwits Web"
##  [38] "Smarp."
##  [39] "Life4mePlus"
##  [40] "HubSpot"
##  [41] "OSdata"
##  [42] "Hocalwire Social Share"
##  [43] "Canva"
##  [44] "United blog to twitter"
##  [45] "dlvr.it"
##  [46] "OptionsProOI"
##  [47] "tickwatcher"
##  [48] "OptionsProVol"
##  [49] "intellinews site integration"
##  [50] "OverBlog Kiwi"
##  [51] "xh_scitech"
##  [52] "Sprinklr"
##  [53] "Foursquare"
##  [54] "Twittimer"
##  [55] "Vinsnobben"
##  [56] "Zoho Social"
##  [57] "CoSchedule"
##  [58] "Jenkers Eng Posting"
##  [59] "Echofon"
##  [60] "Tickeron"
##  [61] "Zapier.com"
##  [62] "Tumblr"
##  [63] "SnapStream TV Search"
##  [64] "SocialNewsDesk"
##  [65] "Troi URL Plug-in"
##  [66] "newsgovhk"
##  [67] "Meltwater Social"
##  [68] "The Healthsite"
##  [69] "SocialBee.io v2"
##  [70] "MTV English News"
##  [71] "Twitter for Advertisers (legacy)"
```

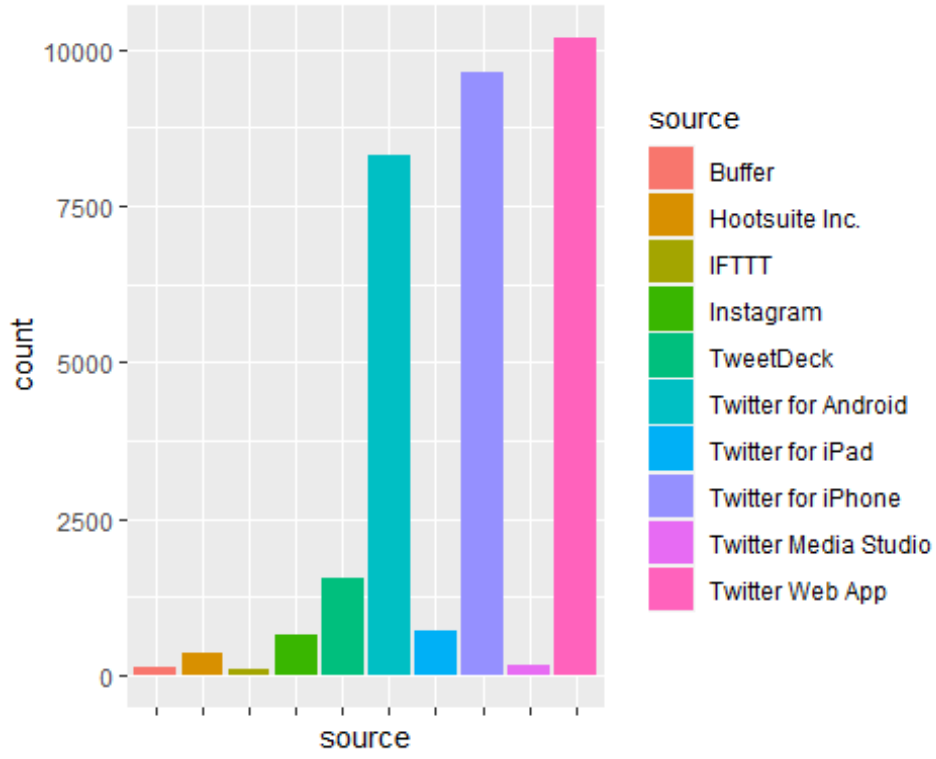
```
##  [72] "PlumeÂ forÂ Android"
##  [73] "Wildmoka"
##  [74] "Salesforce - Social Studio"
##  [75] "Weebly App"
##  [76] "LaterMedia"
##  [77] "AIT News"
##  [78] "eClincher"
##  [79] "HW news english"
##  [80] "FS Poster"
##  [81] "Talon (Classic)"
##  [82] "dailyindia"
##  [83] "presshub_usbot"
##  [84] "TweetCaster for iOS"
##  [85] "Periscope"
##  [86] "Raven Tools"
##  [87] "Article Tweetbot"
##  [88] "iHeartMedia Publishing"
##  [89] "SocialPilot.co"
##  [90] "Publer "
##  [91] "Twitter for Advertisers"
##  [92] "Grabyo"
##  [93] "True Anthem"
##  [94] "Spreaker"
##  [95] "Nelio Content"
##  [96] "AgoraPulse Manager"
##  [97] "btc manager wordpress news"
##  [98] "National Herald"
##  [99] "NDTV News Studio"
## [100] "STOP Imperialism"
## [101] "EveryoneSocial"
## [102] "OptionsMaxPain_Post"
## [103] "ThreadReaderApp"
## [104] "Twitterrific for iOS"
## [105] "Qnary.io"
## [106] "dnh twitter publisher"
## [107] "Twidere X App"
## [108] "Fenix 2"
## [109] "DP-EN - Twitter Auto-Posting"
## [110] "KhuramKTS"
## [111] "SPTK: PutnamDV"
## [112] "Khoros Publishing"
## [113] "Social Reputation"
## [114] "Samrudhi Global"
## [115] "Revive Social App"
## [116] "SEMrush Social Media Tool"
## [117] "OnlyWire App"
## [118] "Oyeyeah"
## [119] "ETRetail.com"
## [120] "Crowdfire App"
## [121] "MarketChameleon.com"
```

```
## [122] "Cloud Campaign"
## [123] "Lightful"
## [124] "El Solucionario App"
## [125] "Constant Contact - Social Posts"
## [126] "aa.com.tr"
## [127] "WP Plugin Dev Com"
## [128] "BLOX CMS"
## [129] "50trends Russia"
## [130] "One News Page (United Kingdom)"
## [131] "godemodebible"
## [132] "Falcon Social Media Management "
## [133] "Woofy Social Media Scheduler"
## [134] "Tweepsmap"
## [135] "Flamingo for Android"
## [136] "TradingView"
## [137] "Scoop.it"
## [138] "djpone"
## [139] "Post Planner Inc."
## [140] "TwInbox"
## [141] "Smart Post App"
## [142] "Janetter Pro for Android"
## [143] "SoCast Digital"
## [144] "SmartNews | ã,¹ãƒžãƒ¼ãƒ\210ãƒ‹ãƒ¥ãƒ¼ã,¹"
## [145] "SOCi - Simplifying Social Media"
## [146] "Mailchimp"
## [147] "EdjNetQuoteFinder"
## [148] "Loomly"
```

Summing all of them in a graph may become tedious and messy. So, for convenience we will analyse the top 10 sources. First let's store it in a new variable 'df1'

```
df1 <- testdata %>%
  group_by(source) %>%
  summarise(count=n()) %>%
  top_n(n=10)

## Selecting by count

df1

## # A tibble: 10 x 2
##    source              count
##    <chr>               <int>
##  1 Buffer                149
##  2 Hootsuite Inc.        356
##  3 IFTTT                  98
##  4 Instagram             665
##  5 TweetDeck            1541
##  6 Twitter for Android  8319
##  7 Twitter for iPad      725
##  8 Twitter for iPhone   9630
```

```
##  9 Twitter Media Studio   162
## 10 Twitter Web App      10201

ggplot(data=df1, aes(x=source, y=count)) + geom_bar(aes(fill=source),
stat='identity') +
  theme(axis.text.x=element_blank())
```



Cleaning Data (Tweets) for Sentiment Analysis:

Convert all text to lower case

```
testdata$text <- iconv(testdata$text,"WINDOWS-1252","UTF-8")
testdata_text <- tolower(testdata$text)
```

Replace blank space

```
testdata_text <- gsub("rt", "", testdata_text)
```

Replace @UserName

```
testdata_text <- gsub("@\\w+", "", testdata_text)
```

Remove punctuation

```
testdata_text <- gsub("[[:punct:]]", "", testdata_text)
```

Remove links

```
testdata_text <- gsub("http\\w+", "", testdata_text)
```

Remove tabs

```
testdata_text <- gsub("[ |\t]{2,}", "", testdata_text)
```

Remove blank spaces at the beginning

```
testdata_text <- gsub("^ ", "", testdata_text)
```

Remove blank spaces at the end

```
testdata_text <- gsub(" $", "", testdata_text)
```

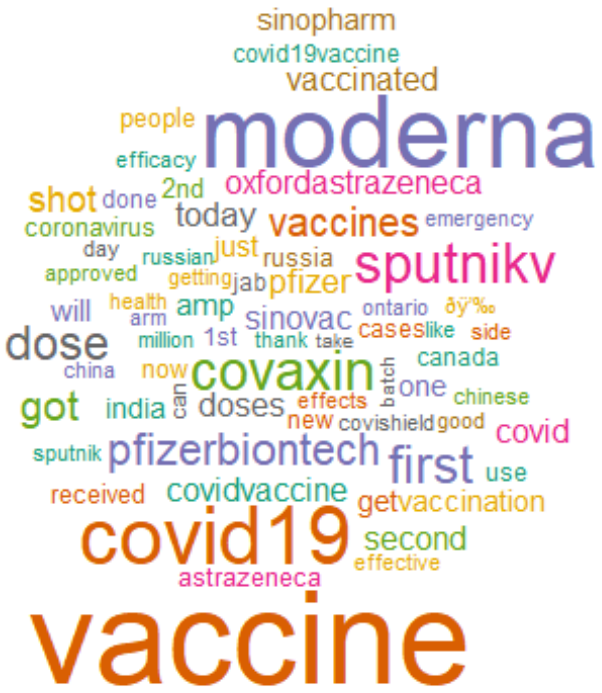
Stop word handling:

Corpus build - remove stop words

```
testdata_text_corpus <- Corpus(VectorSource(testdata_text))
testdata_text_corpus <- tm_map(testdata_text_corpus,
function(x)removeWords(x,stopwords()))

## Warning in tm_map.SimpleCorpus(testdata_text_corpus, function(x)
## removeWords(x, : transformation drops documents
```

Let's display the frequently used words using **word-cloud**

```
wordcloud(testdata_text_corpus,min.freq = 500,colors=brewer.pal(8,
"Dark2"),random.color = TRUE,max.words = 15000)
```



Sentiment Analysis:

Sentiment analysis is typically performed based on a lexicon of sentiment keywords.There are three such sentiment lexicons in **tidytext**:

- The **nrc** lexicon: word and their sentiment category
- The **bing** lexicon: word and their polarity (negative or positive)
- The **ann** lexicon: word and their numeric sentiment score

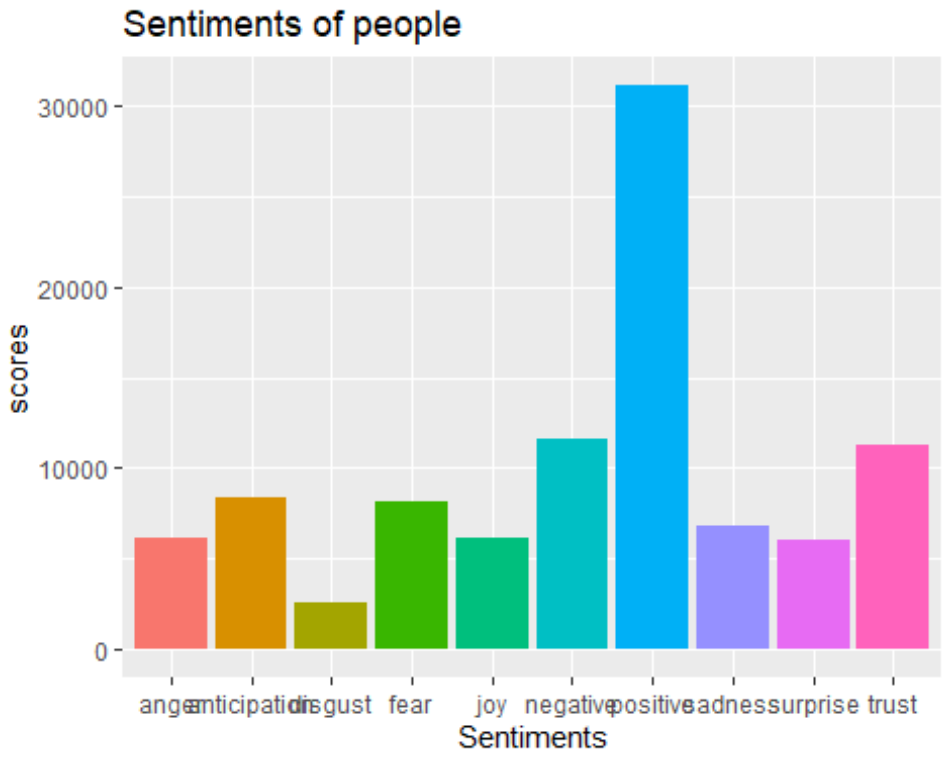
```
testdata_text_sent<-get_nrc_sentiment((testdata_text))
```

Now, Let's calculate the total score for each sentiment

```
testdata_text_sent_score<-data.frame(colSums(testdata_text_sent[,]))

names(testdata_text_sent_score)<-"Score"
testdata_text_sent_score<-
cbind("sentiment"=rownames(testdata_text_sent_score),testdata_text_sent_score
)
rownames(testdata_text_sent_score)<-NULL
```

Now, Let's plot the sentiments with scores

```
ggplot(data=testdata_text_sent_score,aes(x=sentiment,y=Score))+geom_bar(aes(f
ill=sentiment),stat = "identity")+
  theme(legend.position="none")+
  xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people ")
```



Let's remove positive , negative score

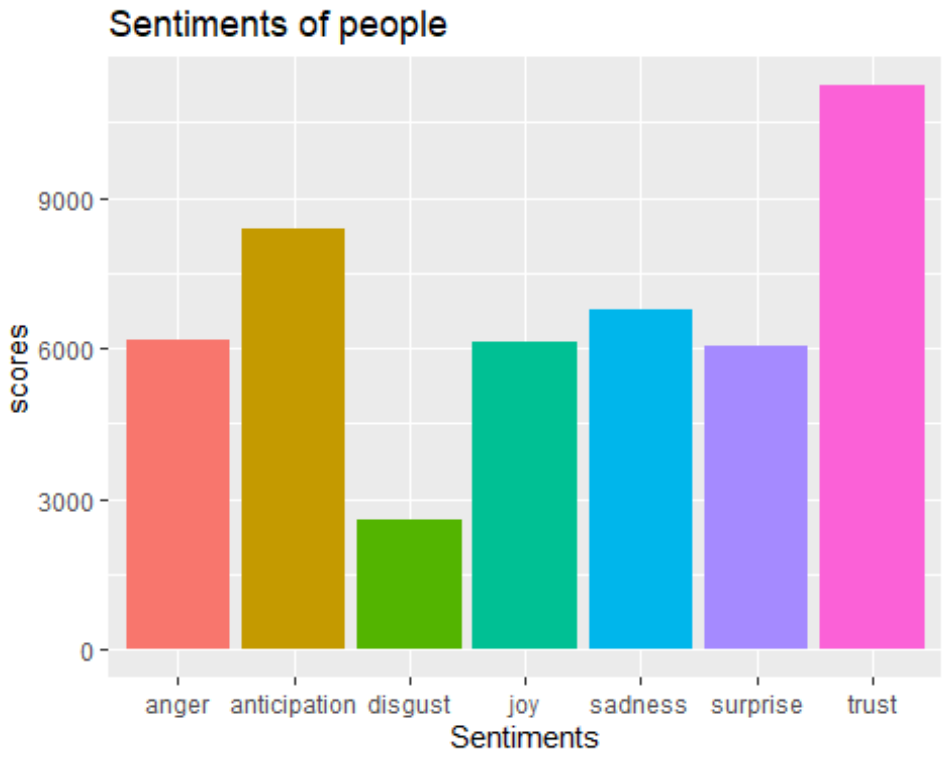
```
testdata_text_sent<-get_nrc_sentiment((testdata_text))

testdata_text_sent_no_pos_neg<-
select(testdata_text_sent,anger,anticipation,disgust,joy,sadness,surprise,tru
st)
```

Now, Let's calculate the total score for each sentiment

```
testdata_text_sent_no_pos_neg<-
data.frame(colSums(testdata_text_sent_no_pos_neg[,]))

names(testdata_text_sent_no_pos_neg)<-"Score"
testdata_text_sent_no_pos_neg<-
cbind("sentiment"=rownames(testdata_text_sent_no_pos_neg),testdata_text_sent_
no_pos_neg)
rownames(testdata_text_sent_no_pos_neg)<-NULL
```

Now, Let's plot the sentiments with scores

```
ggplot(data=testdata_text_sent_no_pos_neg,aes(x=sentiment,y=Score))+geom_bar(
aes(fill=sentiment),stat = "identity")+
  theme(legend.position="none")+
  xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people ")
```



Conclusion:

From the above graph, we can conclude that, people are showing trust and overall positive emotions for the covid vaccine . The anticipation of people is high.