DUBLIN BUSINESS SCHOOL

SUMMATIVE ASSESSMENT - EDA

B89IT106 DATA VISUALISATION

TERRI HOARE

Nishad Abdul latheef
Student ID: 10382242
Date: 8-10-2018

Contents

Introduction 3
Dataset 4
Tableau Visualisation 6
R Visualisation 14
Conclusion 23
References 24

Introduction

Data visualization is defined as an attempt to aid people understand and explore the relevance of data by converting it into a visual form. The target of our assignment is to apply visualisation techniques to the output dataset of Sentimental Analysis of Twitter data.

Tools used from visualisation:

1. **RStudio:**

RStudio is a free and open-source integrated development environment (IDE) for R.

2. **Tableau**

Tableau is a data visualization tool used for business intelligence. With the help of Tableau, one can build a very interactive visual within minutes.

Dataset

The dataset used is the output of the sentiment analysis of twitter data, which is a technique that extracts the emotions of each tweet. We used the *twitter search* operator (Fig 1) to extract tweets about a specific topic (here, names of different political personalities were used).

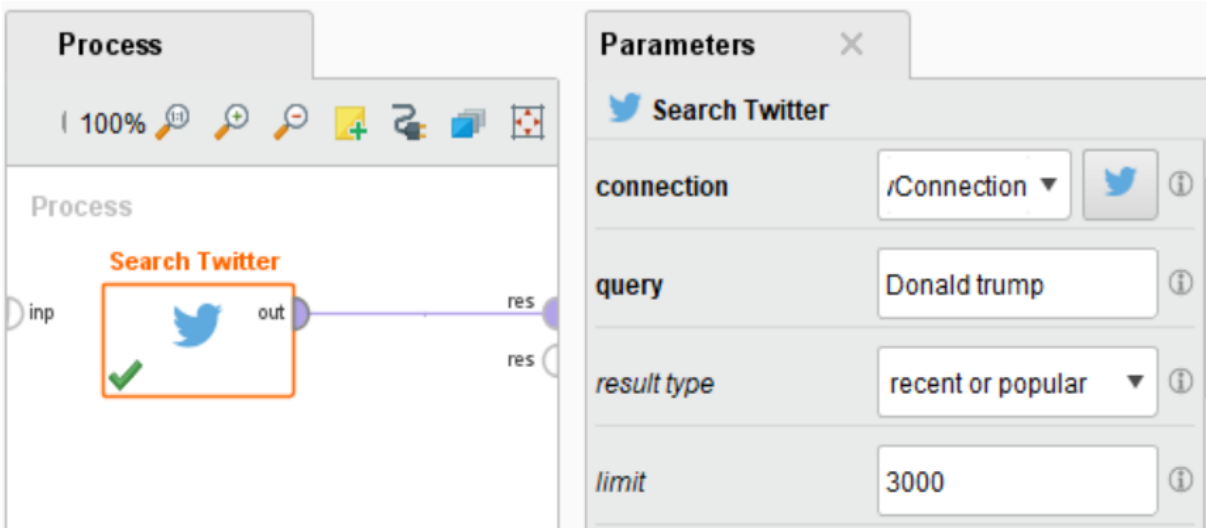


Fig 1

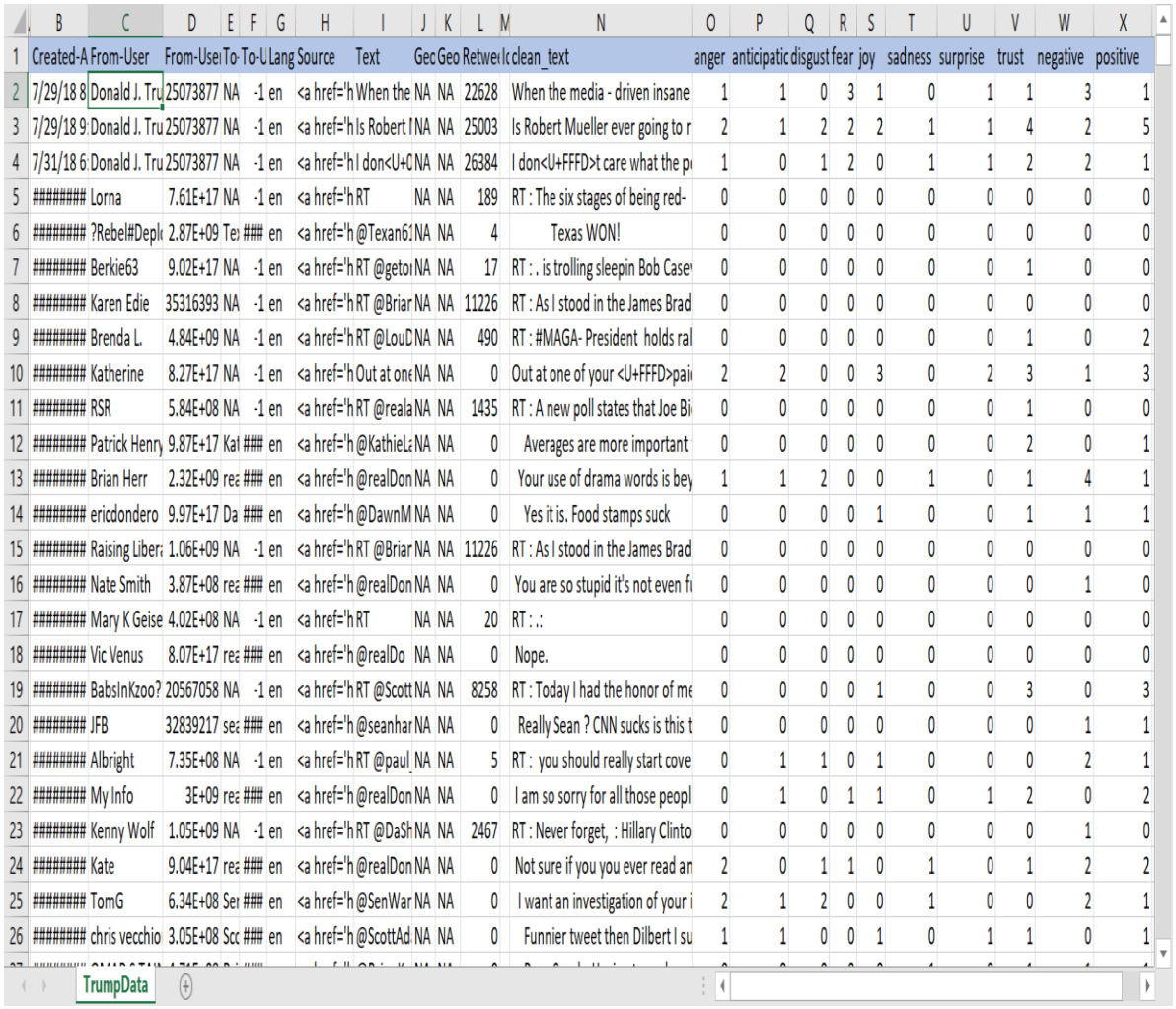
The extracted data was processed in R for Sentiment Analysis with the given code:

```
## Importing the tweets from dataset
tweet <- TrunpTweets
class(tweet)


## Cleaning text
tweet$clean_text <- str_replace_all(tweet$Text, "@\\w+", "")


## Sentiment analysis
Sentiment <- get_nrc_sentiment(tweet$clean_text)
tweets_senti <- cbind(tweet, Sentiment)
View(tweets_senti)
```

The final dataset (Fig 2) that was used for data visualisation contains 3000 rows and 22 attributes that includes 8 emotions, 2 polarity, tweet texts, etc.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Created-A	From-User	From-Use	To-	To-U	Lang	Source	Text	Geo	Geo	Retwe	Ic	clean_text	anger	anticipatic	disgust	fear	joy	sadness	surprise	trust	negative	positive
2	7/29/18 8	Donald J. Tru	25073877	NA	-1	en	<a href='h	When the	NA	NA	22628		When the media - driven insane	1	1	0	3	1	0	1	1	3	1
3	7/29/18 9	Donald J. Tru	25073877	NA	-1	en	<a href='h	Is Robert I	NA	NA	25003		Is Robert Mueller ever going to r	2	1	2	2	2	1	1	4	2	5
4	7/31/18 6	Donald J. Tru	25073877	NA	-1	en	<a href='h	I don<U+0	NA	NA	26384		I don<U+FFFD>t care what the p	1	0	1	2	0	1	1	2	2	1
5	########	Lorna	7.61E+17	NA	-1	en	<a href='h	RT	NA	NA	189		RT : The six stages of being red-	0	0	0	0	0	0	0	0	0	0
6	########	?Rebel#Depl	2.87E+09	Tex	###	en	<a href='h	@Texan61	NA	NA	4		Texas WON!	0	0	0	0	0	0	0	0	0	0
7	########	Berkie63	9.02E+17	NA	-1	en	<a href='h	RT @getor	NA	NA	17		RT : . is trolling sleepin Bob Case	0	0	0	0	0	0	0	1	0	0
8	########	Karen Edie	35316393	NA	-1	en	<a href='h	RT @Briar	NA	NA	11226		RT : As I stood in the James Brad	0	0	0	0	0	0	0	0	0	0
9	########	Brenda L.	4.84E+09	NA	-1	en	<a href='h	RT @LouD	NA	NA	490		RT : #MAGA- President holds ral	0	0	0	0	0	0	0	1	0	2
10	########	Katherine	8.27E+17	NA	-1	en	<a href='h	Out at one	NA	NA	0		Out at one of your <U+FFFD>pai	2	2	0	0	3	0	2	3	1	3
11	########	RSR	5.84E+08	NA	-1	en	<a href='h	RT @reala	NA	NA	1435		RT : A new poll states that Joe Bi	0	0	0	0	0	0	0	1	0	0
12	########	Patrick Henry	9.87E+17	Kat	###	en	<a href='h	@KathieL	NA	NA	0		Averages are more important	0	0	0	0	0	0	0	2	0	1
13	########	Brian Herr	2.32E+09	rea	###	en	<a href='h	@realDon	NA	NA	0		Your use of drama words is bey	1	1	2	0	0	1	0	1	4	1
14	########	ericdondero	9.97E+17	Da	###	en	<a href='h	@DawnM	NA	NA	0		Yes it is. Food stamps suck	0	0	0	0	1	0	0	1	1	1
15	########	Raising Liber	1.06E+09	NA	-1	en	<a href='h	RT @Briar	NA	NA	11226		RT : As I stood in the James Brad	0	0	0	0	0	0	0	0	0	0
16	########	Nate Smith	3.87E+08	rea	###	en	<a href='h	@realDon	NA	NA	0		You are so stupid it's not even f	0	0	0	0	0	0	0	0	1	0
17	########	Mary K Geise	4.02E+08	NA	-1	en	<a href='h	RT	NA	NA	20		RT : .:	0	0	0	0	0	0	0	0	0	0
18	########	Vic Venus	8.07E+17	rea	###	en	<a href='h	@realDo	NA	NA	0		Nope.	0	0	0	0	0	0	0	0	0	0
19	########	BabsInKzoo?	20567058	NA	-1	en	<a href='h	RT @Scott	NA	NA	8258		RT : Today I had the honor of me	0	0	0	0	1	0	0	3	0	3
20	########	JFB	32839217	sea	###	en	<a href='h	@seanhar	NA	NA	0		Really Sean ? CNN sucks is this t	0	0	0	0	0	0	0	0	1	1
21	########	Albright	7.35E+08	NA	-1	en	<a href='h	RT @paul_	NA	NA	5		RT : you should really start cove	0	1	1	0	1	0	0	0	2	1
22	########	My Info	3E+09	rea	###	en	<a href='h	@realDon	NA	NA	0		I am so sorry for all those peopl	0	1	0	1	1	0	1	2	0	2
23	########	Kenny Wolf	1.05E+09	NA	-1	en	<a href='h	RT @DaSh	NA	NA	2467		RT : Never forget, : Hillary Clinto	0	0	0	0	0	0	0	0	1	0
24	########	Kate	9.04E+17	rea	###	en	<a href='h	@realDon	NA	NA	0		Not sure if you you ever read an	2	0	1	1	0	1	0	1	2	2
25	########	TomG	6.34E+08	Ser	###	en	<a href='h	@SenWar	NA	NA	0		I want an investigation of your i	2	1	2	0	0	1	0	0	2	1
26	########	chris vecchio	3.05E+08	Sco	###	en	<a href='h	@ScottAd	NA	NA	0		Funnier tweet then Dilbert I su	1	1	0	0	1	0	1	1	0	1

TrumpData

Fig 2

Tableau Visualisation

We chose Tableau as one of the visualisation tools because of the following reasons:

1. **Tableau clearly and beautifully visualizes your data**

   Tableau can tell stories with simple visualizations, making it easy for your clients to understand.

2. **Tableau is easy to use**

   Tableau makes it easy for users to use on a regular basis. The desktop application is a simple authoring tool for creating your

3. **Tableau has an excellent user-experience**

   The familiar drag-and-drop interface makes it similar to Excel and, again, the visualization options are abundant.

3. **Tableau can handle large amounts of data**

   Tableau has the ability to produce reports on extremely large sets of data without drastically affecting network performance.

The dataset was loaded into tableau for processing and visualising of data.

1. **Emotion Count**

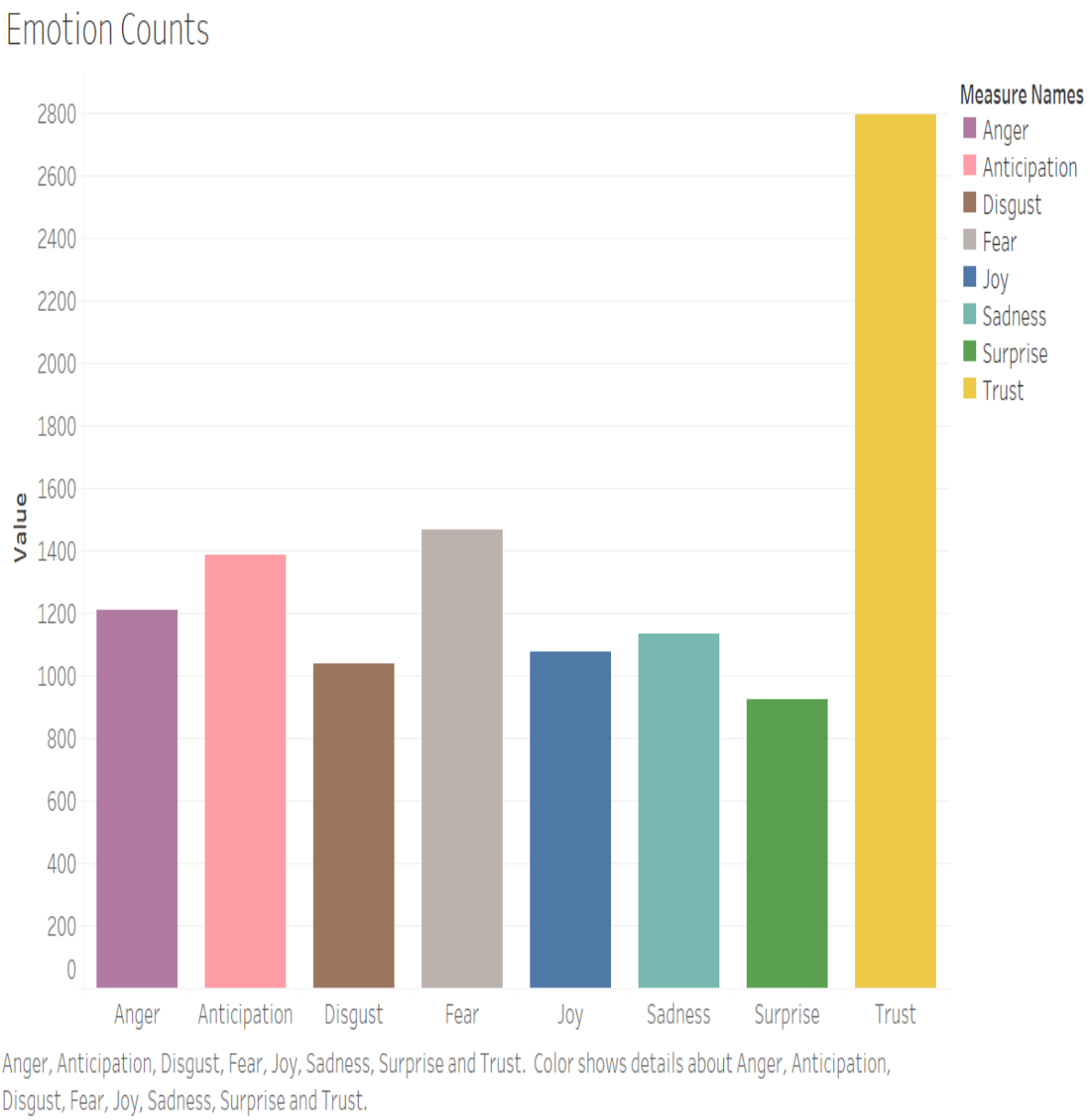


Fig 3

The emotion count of the tweets mentioning **Donald Trump**'s name was plotted (Fig 3). From the plot it is visible that the most common emotion on that period was **Trust.** On the other hand, the least detected emotion is **Surprise.**

2. Polarity Count

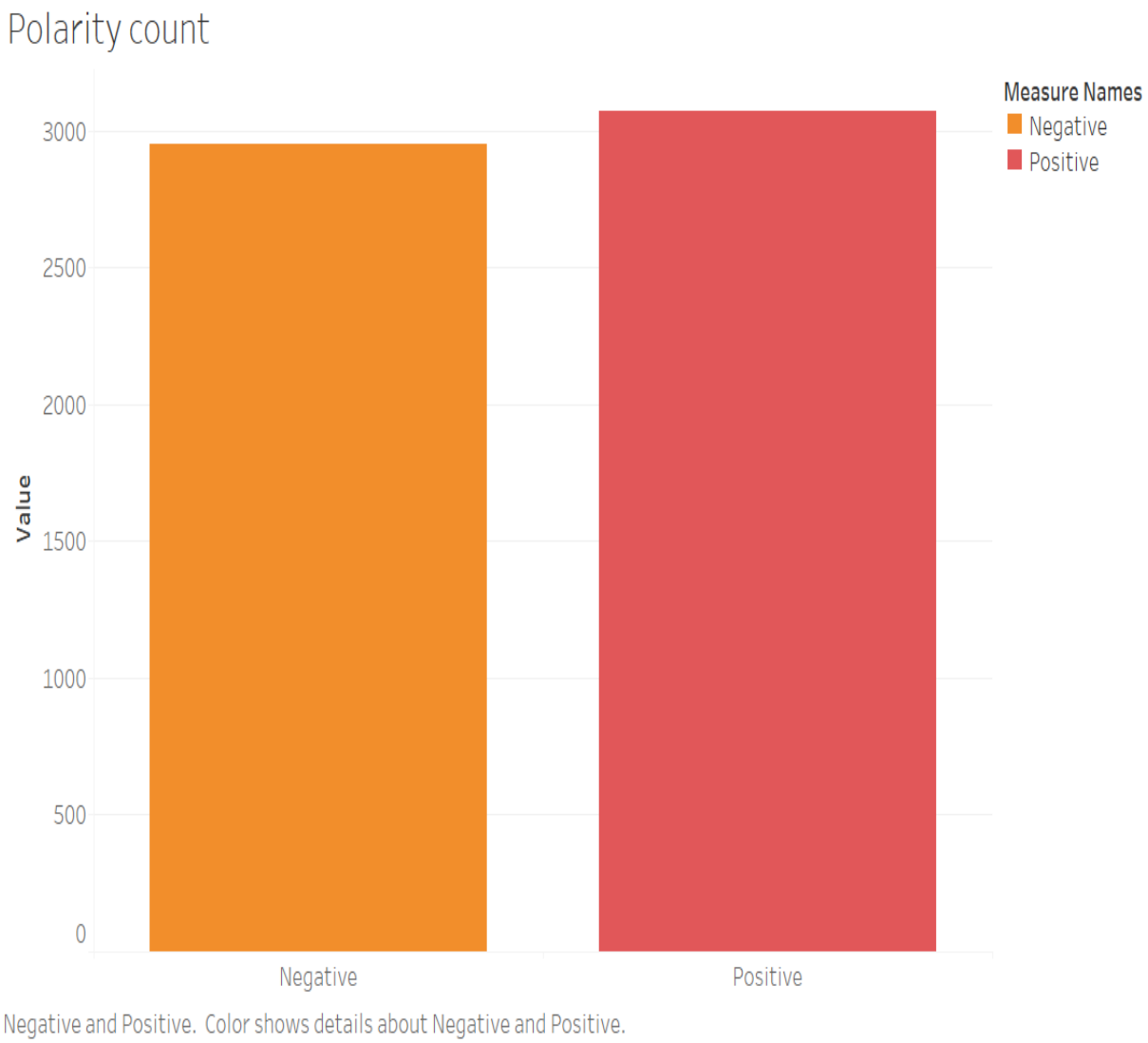


Fig 4

The polarity count of the tweets mentioning **Donald Trump**'s name was plotted (Fig 4). The polarity is calculated as the score of the emotions. In a given period, the **Donald Trump** gets more positive emotions than negative. The results show that the support for Trump is still alive.

3. Timeline

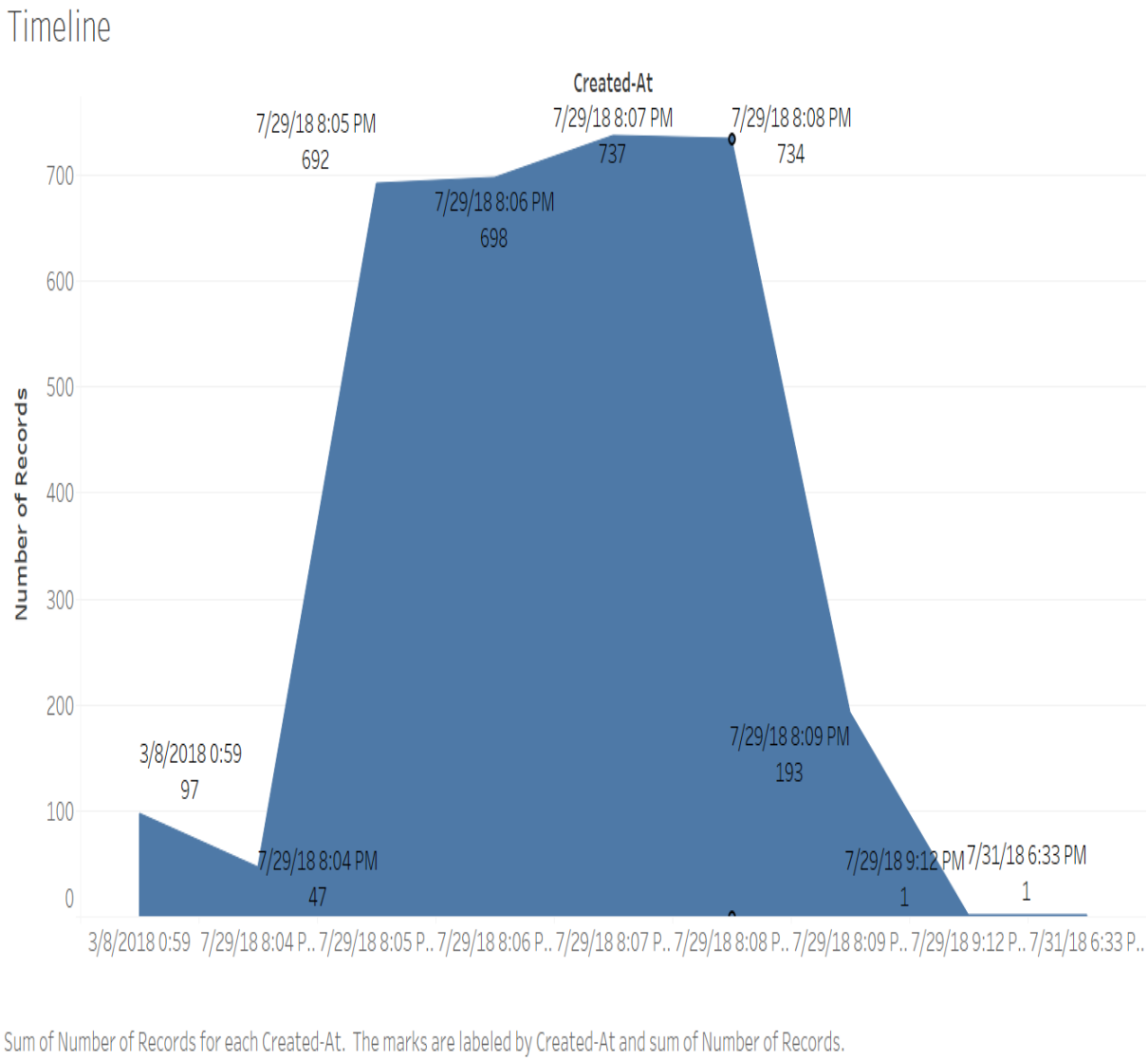


Fig 5

The Timeline of the tweets mentioning **Donald Trump**'s name was plotted (Fig 5). The highest number of tweets were created at **29-07-2018 8:07 PM** with the total count of **737.** The least number of tweets were created at **31-07-2018 6:33 PM** & **29-07-2018 9:12 PM** with the total count of **1.**

4. Users

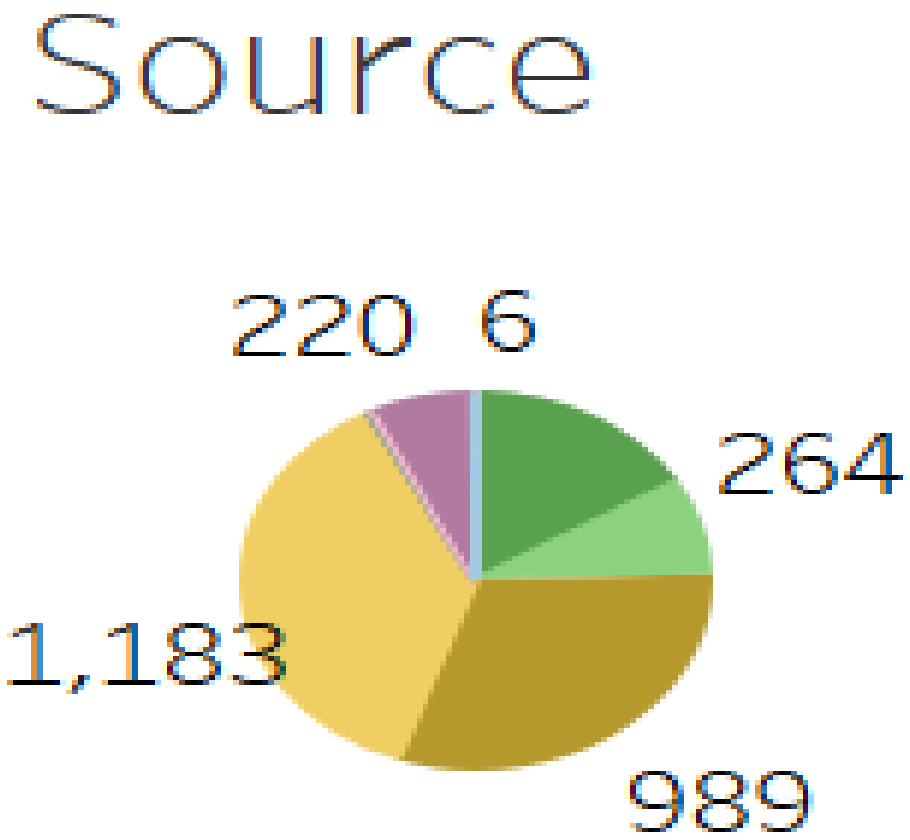
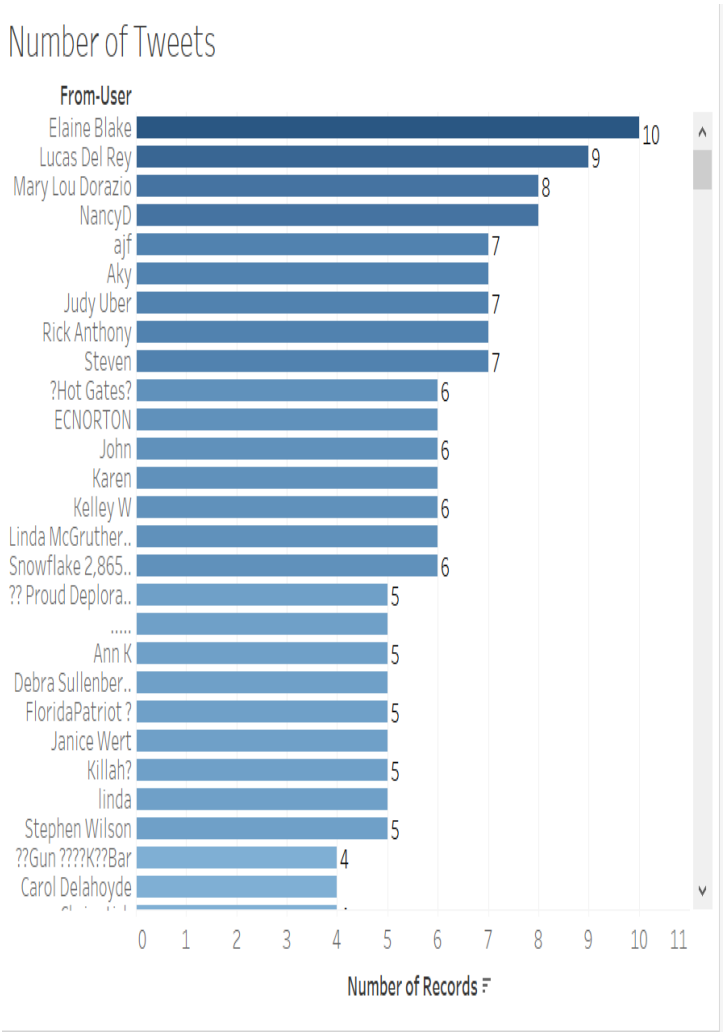


Fig 6

The Pie Chart of the source of tweets mentioning **Donald Trump**'s name was plotted (Fig 6). The chart shows that highest number of from a specific source is from **Twitter for iPhone** with **1,183** users followed by **Twitter for Android** with **989** users.

5. Tweets and Retweets



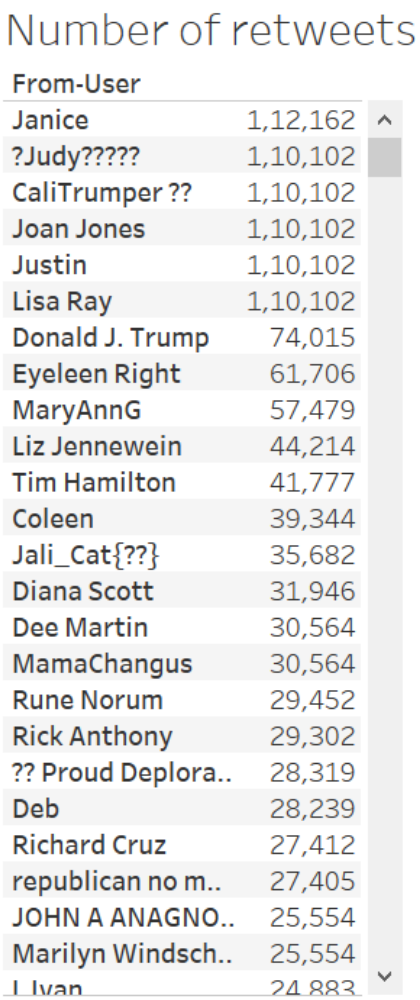


Fig 7

The data of the number of tweets and retweets mentioning **Donald Trump**'s name was plotted (Fig 7). The greatest number of tweets about Trump was tweeted by **Elaine Blake** with **10** tweets. On the other hand, **Janice** had the most retweets with **112,162.**

6. Dashboards

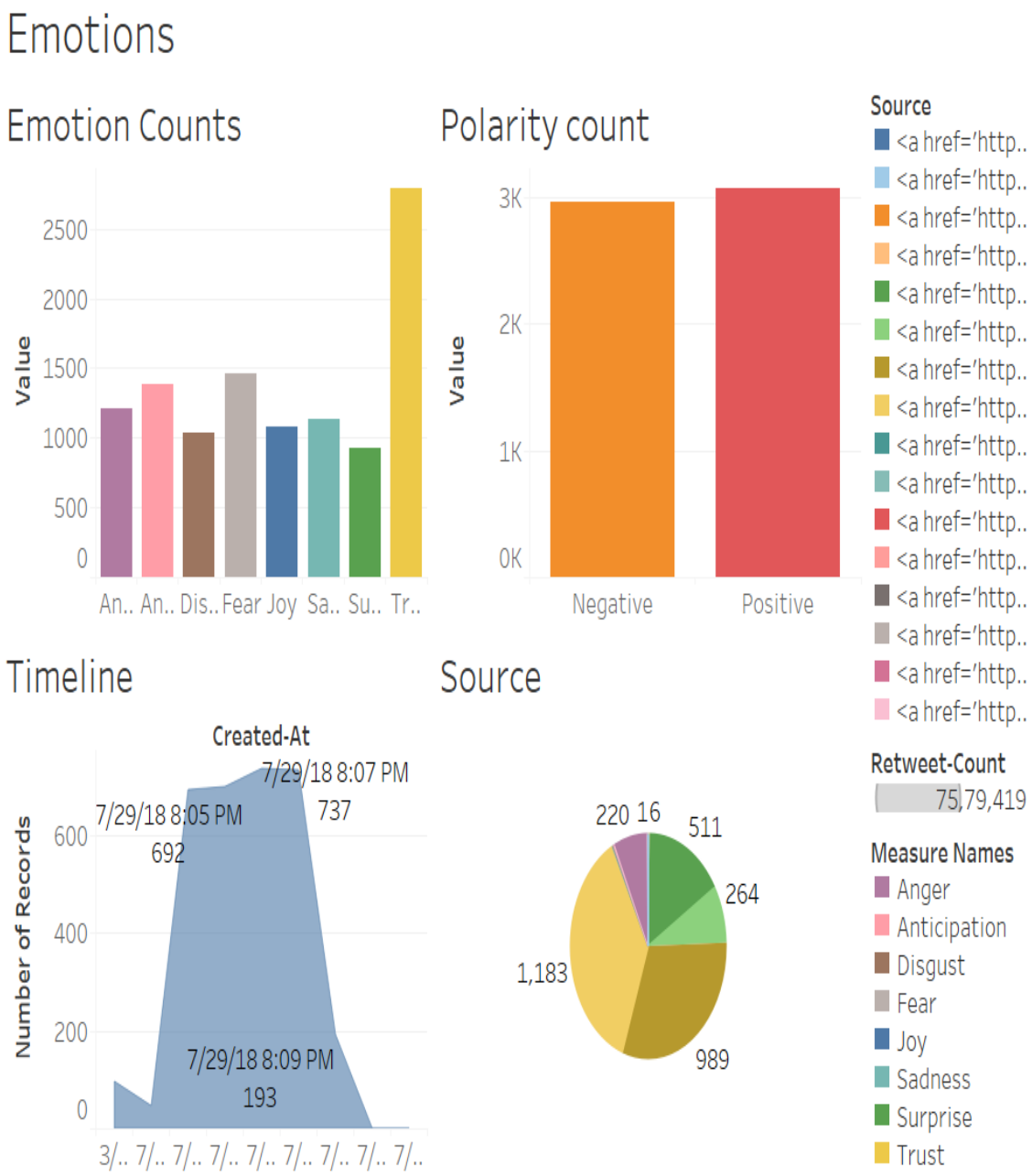


Fig 8

The first dashboard(Fig 8) was built with the worksheets named as Emotions Counts, Polarity count, Timeline and Source.

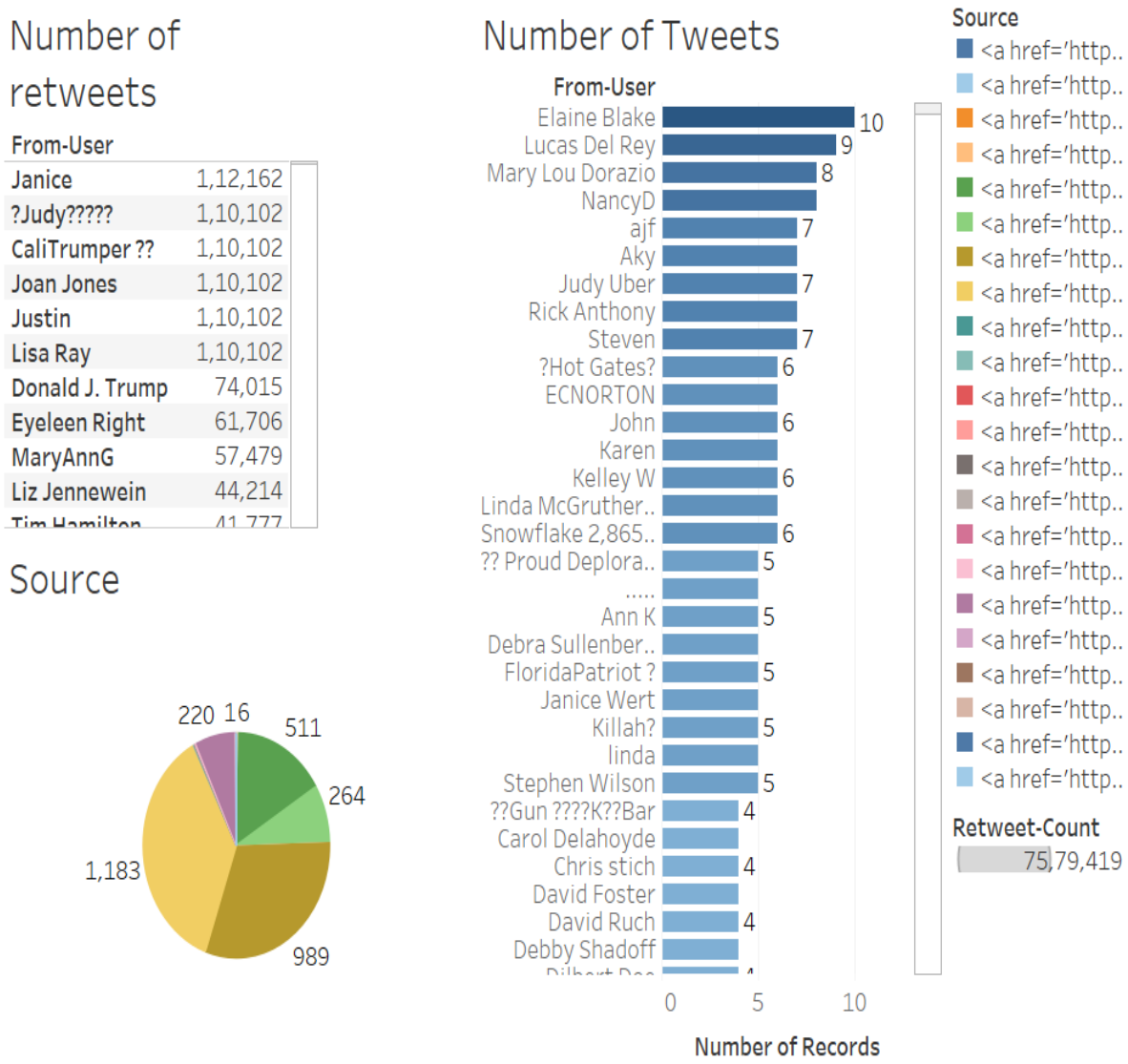


Fig 9

The first dashboard (Fig 9) was built with the worksheets named as Source, Number of Tweets and Number of Retweets.

R Visualisation

We chose RStudio as one of the visualisation tools and dataset with similar format with the key word **“Narendra Modi”** was used for visualisation.

The libraries used for visualisation are:

1. **ggplot2**

   ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

2. **scales**

   One of the most difficult parts of any graphics package is scaling, converting from data values to perceptual properties. The inverse of scaling, making guides (legends and axes) that can be used to read the graph, is often even harder. The idea of the scales package is to implement scales in a way that is graphics system agnostic, so that everyone can benefit by pooling knowledge and resources about this tricky topic.

3. **wordcloud**

   Plot a cloud of words shared across documents.

4. **RColorBrewer**

   Creates nice looking colour palettes especially for thematic maps.

1. **Tweets by month**

The dataset was loaded into the Rstudio and the following code was executed:

```
##Plotting the tweets by month

ggplot(data = tweet, aes(x = month

(`Created.At`, label = TRUE))) +

  geom_bar(aes(fill = ..count..)) +

  theme(legend.position = "none") +

  xlab("Month") + ylab("Number of tweets") +

  scale_fill_gradient(low = "blue", high = "red")
```

This gave the following visualisation of tweets based on months (Fig 10).

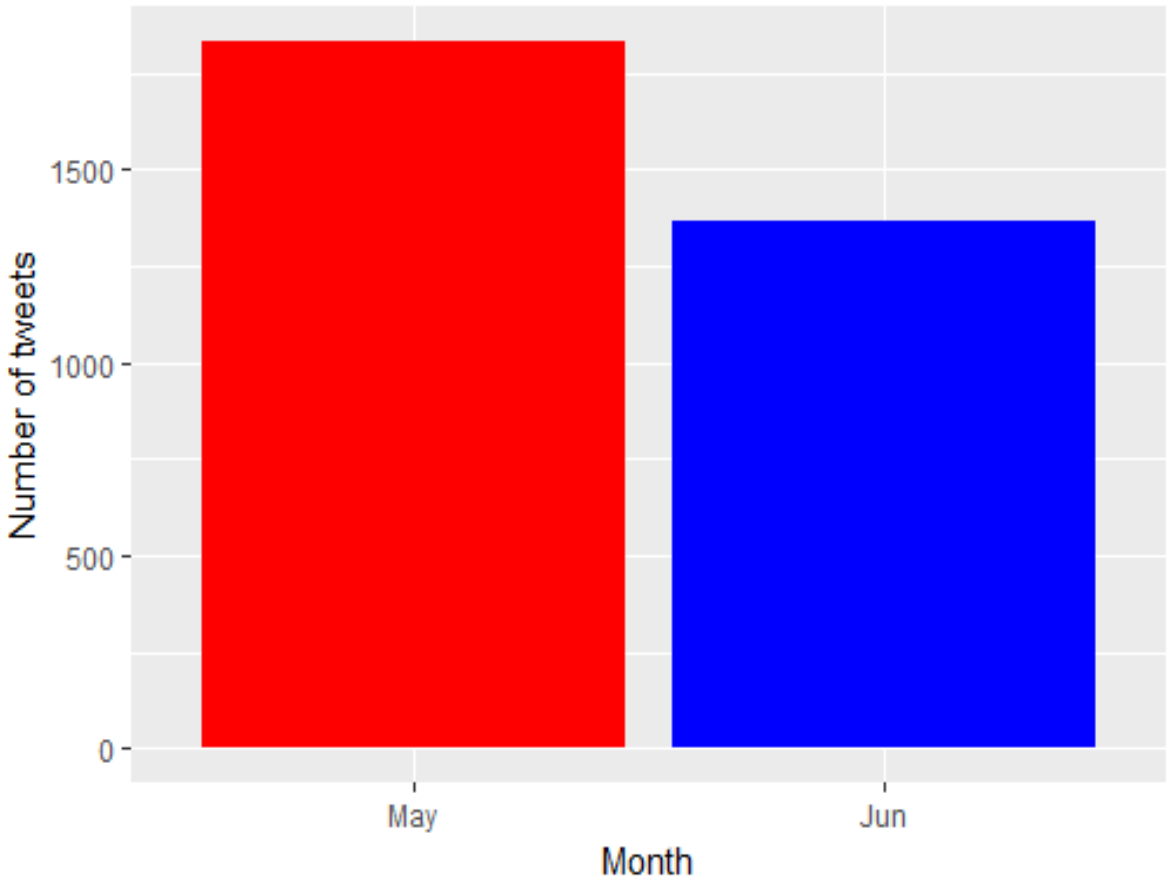


Fig 10

The tweets recorded on the month of May was more compared to the tweets on June.

2. Tweets by Time

The following Rcode was executed and the output was obtained(Fig11):

```
## Grouping tweets by time
tweet$timeonly      <-      as.numeric(tweet$`Created.At`      -
trunc(tweet$`Created.At`, "days"))
class(tweet$timeonly) <- "POSIXct"
## Plotting by time
ggplot(data = tweet, aes(x = timeonly)) +
  geom_histogram(aes(fill = ..count..)) +
  theme(legend.position = "none") +
  xlab("Time") + ylab("Number of tweets") +
  scale_x_datetime(breaks = date_breaks("3 hours"),
                   labels = date_format("%H:00")) +
  scale_fill_gradient(low = "blue", high = "red")
```

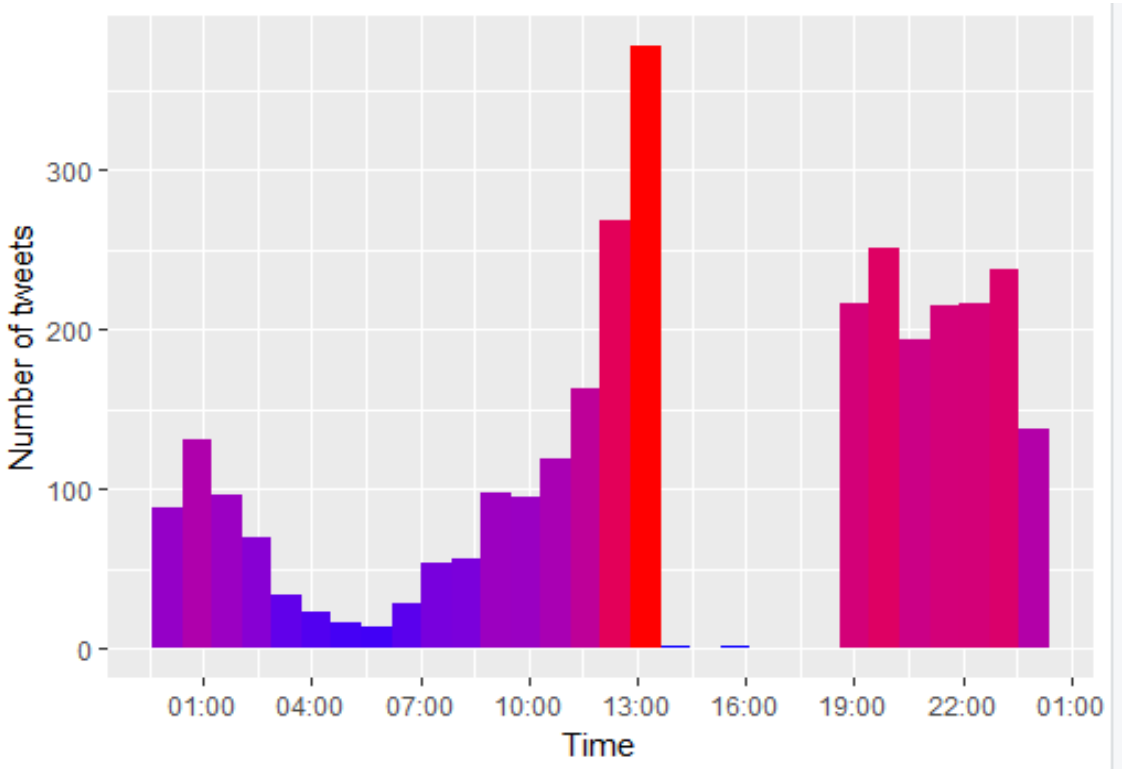


Fig 11

The plot shows that highest number of tweets was recorded at **1:00 PM** and least during **2:00 PM to 6:00 PM.**

3. Word Clouds

The following R code was executed.

```
## CLeaning the text
nohandles <- str_replace_all(tweet$Text, "@\\w+", "") ##
Removing symbols

## Creating the corpus
wordCorpus <- Corpus(VectorSource(nohandles))
wordCorpus <- tm_map(wordCorpus, removePunctuation)
wordCorpus <- tm_map(wordCorpus, content_transformer(tolower))
#converted to lower case
wordCorpus <- tm_map(wordCorpus, removeWords,
stopwords("english")) #remove stopwords
wordCorpus <- tm_map(wordCorpus, removeWords, c("amp", "2yo",
"3yo", "4yo"))
wordCorpus <- tm_map(wordCorpus, stripWhitespace)

## Plotting th corpus
pal <- brewer.pal(9,"YlGnBu")
pal <- pal[-(1:4)]
set.seed(123)

## Creating a wordcloud
wordcloud(words = wordCorpus, scale=c(10,0.3), max.words=100,
random.order=FALSE,
          rot.per=0.35, use.r.layout=FALSE, colors=pal)
```

The output of the code (fig 12) showed the word-cloud of the most occurring words in the tweets.

Fig 12

4. Emotion count

Similar to tableau visualisation, bar charts of emotion count where plotted with the following RCode.

```
## Total of the sentiment weights
sentimentTotals <- data.frame(colSums(tweets_senti[,c(15:22)]))
names(sentimentTotals) <- "count"

## Joining the datasets
sentimentTotals <- cbind("sentiment" =
rownames(sentimentTotals), sentimentTotals)
rownames(sentimentTotals) <- NULL

## Plotting the sentiment scores
ggplot(data = sentimentTotals, aes(x = sentiment, y = count)) +
  geom_bar(aes(fill = sentiment), stat = "identity") +
  theme(legend.position = "none") +
  xlab("Sentiment") + ylab("Total Count") + ggtitle("Total
Sentiment Score for All Tweets")
```

The executed r code gave the output as (Fig 13):

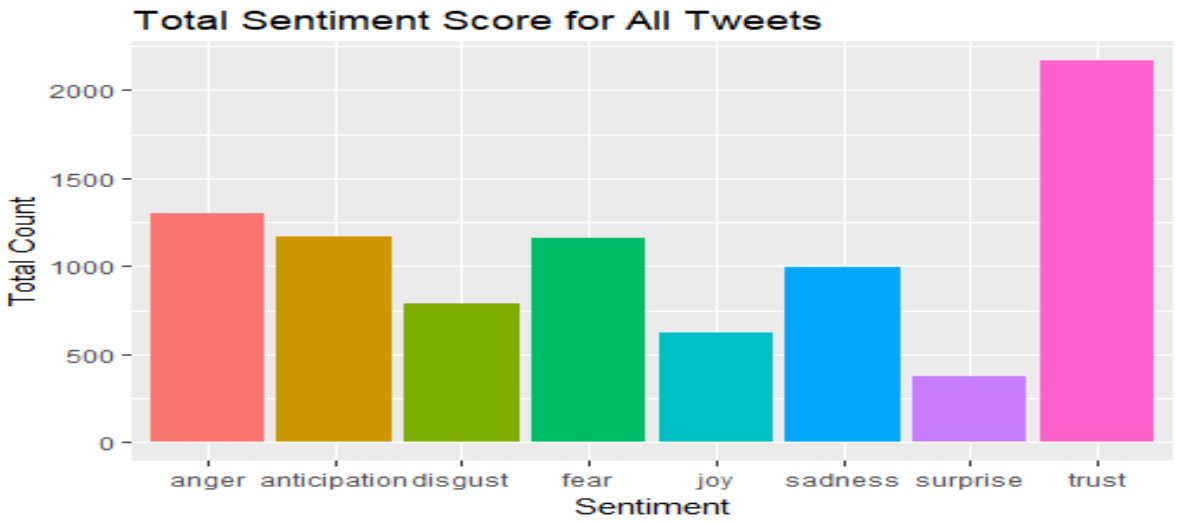


Fig 12

From the graph, we can see that, most of the tweeters showed trust as their emotion.

5. **Polarity**

The polarities of tweets were plotted using the R code:

```
## Positive and negative tweets

posnegtime <- tweets_senti %>%

  group_by(created = cut(`Created.At`, breaks="10 hour")) %>%

  summarise(negative = mean(negative),

            positive = mean(positive)) %>% melt
```

The output obtained looked like (Fig 14):

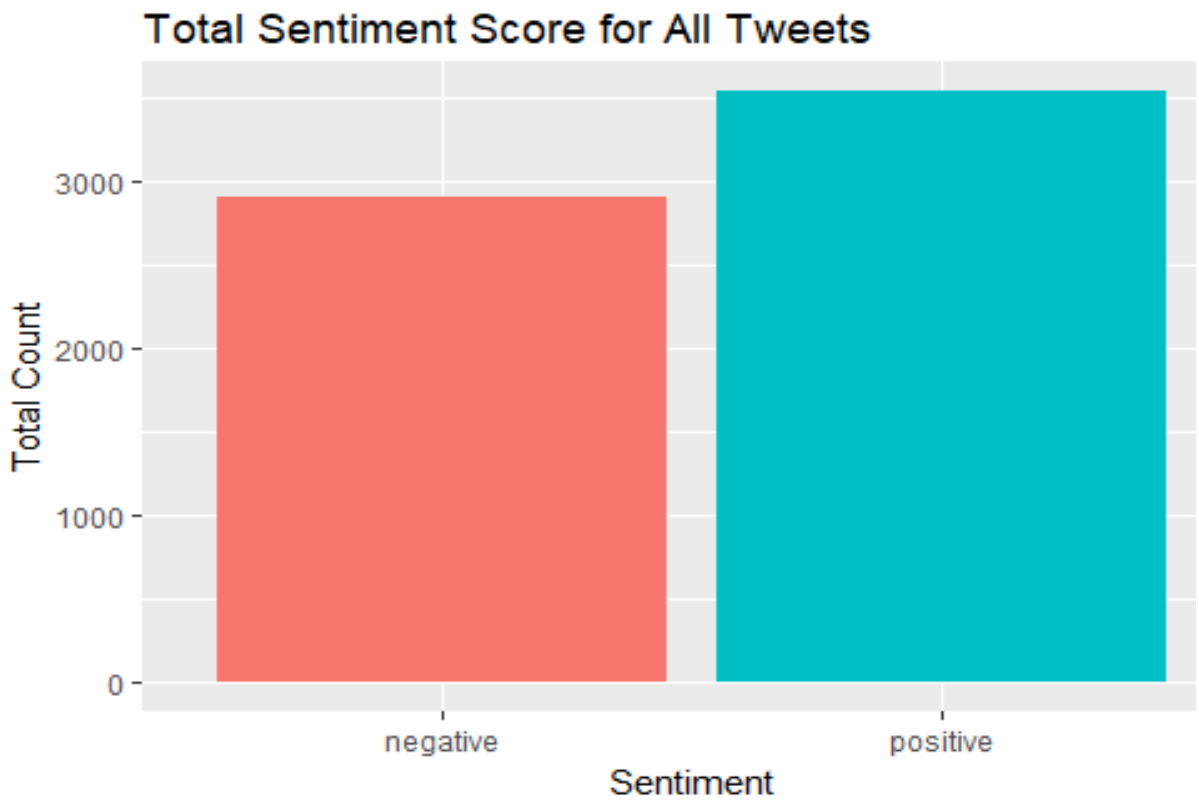


Fig 14

6. Sentiment with Time

The variation of sentiment with time was plotted using the code:

```
## Plotting the tweets
ggplot(data = posnegtime, aes(x = as.Date(timestamp), y =
meanvalue, group = sentiment)) +
  geom_line(size  =  1.5,  alpha  =  0.7,  aes(color  =
sentiment)) +
  geom_point(size = 0.3) +
  ylim(0, NA) +
  scale_colour_manual(values = c("green", "red")) +
  theme(legend.title=element_blank(),   axis.title.x   =
element_blank()) +
  scale_x_time(breaks = waiver(), minor_breaks = waiver(),
               labels = date_format("%h-%m")) +
  ylab("Average sentiment score") +
  ggtitle("Sentiment Over Time")

## Grouping by emotion

tweets_senti$day <- wday(tweets_senti$`Created.At`, label
= TRUE)
dailysentiment <- tweets_senti %>% group_by(day) %>%
  summarise(anger = mean(anger),
            anticipation = mean(anticipation),
            disgust = mean(disgust),
            fear = mean(fear),
            joy = mean(joy),
            sadness = mean(sadness),
            surprise = mean(surprise),
            trust = mean(trust)) %>% melt

## Variation in sentiment by day
names(dailysentiment)    <-    c("day",    "sentiment",
"meanvalue")

## Plotting the emotion variation by day
ggplot(data = dailysentiment, aes(x = day, y = meanvalue,
group = sentiment)) +
  geom_line(size  =  2.5,  alpha  =  0.7,  aes(color  =
sentiment)) +
  geom_point(size = 0.5) +
  ylim(0, NA) +
  theme(legend.title=element_blank(),   axis.title.x   =
element_blank()) +
  ylab("Average sentiment score") +
  ggtitle("Sentiment During the Year")
```

The outputs (Fig 15 & Fig16) were plotted.

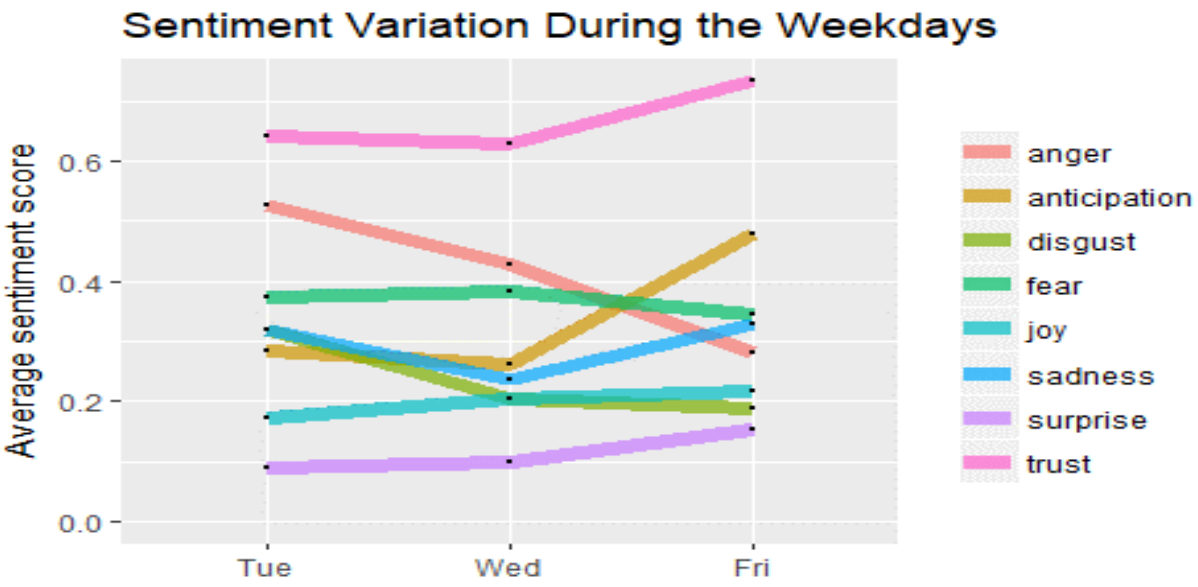


Fig 15

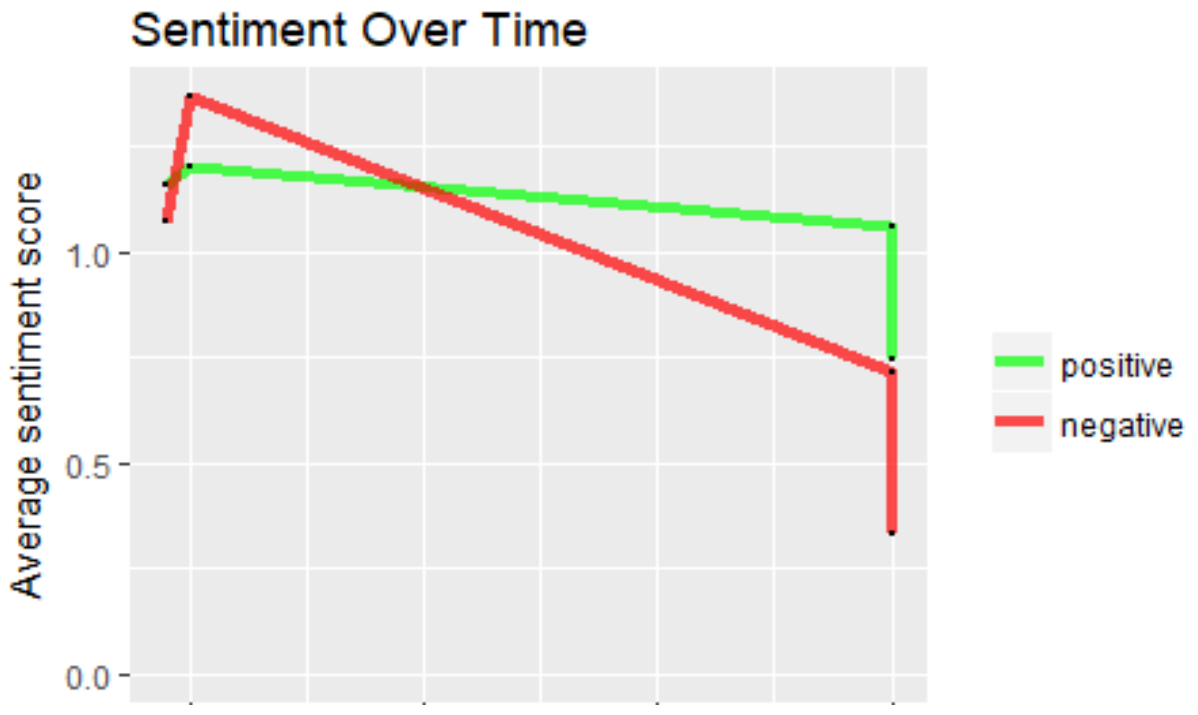


Fig 16

From the plots, it is clear that the number negative tweets have reduced and the number of positive tweets has increased along the week.

Conclusion

After exploring the various visualisation techniques, I have observed the following about the two platforms:

- Tableau:

  It is a very interactive and user-friendly tool in which visualisation could be applied rapidly due to its software designs. The recommendation of the techniques that you get once you load the data is a very useful feature. The choice of building a Story from dashboard and dashboard from worksheets helps the visualisation process more fun.

- R Studio:

  The usage of ggplot package in R makes the visualisation attractive for the user. The demerit of taking a lot of time to build the visualisation compared to tableau can be compensated with the interesting output that it leads to. The added advantage of content manipulation is definitely something worth spending time for.

References

Kaggle.com. (2018). *Kaggle: Your Home for Data Science*. [online] Available at: https://www.kaggle.com/ [Accessed 23 Jul. 2018].

Rdocumentation.org. (2018). *R Documentation and manuals | R Documentation*. [online] Available at: https://www.rdocumentation.org/ [Accessed 15 Jul. 2018].

Tableau Software. (2018). *Tableau Software*. [online] Available at: https://www.tableau.com/ [Accessed 1 Aug. 2018].

Wikipedia.org. (2018). *Wikipedia*. [online] Available at: https://www.wikipedia.org/ [Accessed 4 Aug. 2018].