QUESTION A1

```
$ pwd
/c/Users/HP PROBOOK

$ cd /c/Users/HP\ PROBOOK/Documents/FIT1043/A3

$ ls
'FIT1043 A3 Specification- S2 2024.pdf'   FIT1043_Dataset

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ ls -lh FIT1043_Dataset.gz
-rw-r--r-- 1 HP PROBOOK 197609 74M Oct 16 21:29 FIT1043_Dataset.gz

$ ls -lh FIT1043_Dataset.gz
-rw-r--r-- 1 HP PROBOOK 197609 74M Oct 14 15:26 FIT1043_Dataset.gz
```

74 Megabytes in FIT1043_Dataset

QUESTION A1.2

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ head -n 5 FIT1043_Dataset
0,1467810672,Mon Apr 06 22:19:49 PDT 2009,NO_QUERY,scotthamilton,is upset that h
e can't update his Facebook by texting it... and might cry as a result  School t
oday also. Blah!
0,1467810917,Mon Apr 06 22:19:53 PDT 2009,NO_QUERY,mattycus,@Kenichan I dived ma
ny times for the ball. Managed to save 50%  The rest go out of bounds
0,1467811184,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,ElleCTF,my whole body feels i
tchy and like its on fire
0,1467811193,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,Karoli,@nationwideclass no it
's not behaving at all. i'm mad. why am i here? because I can't see you all over
 there.
0,1467811372,Mon Apr 06 22:20:00 PDT 2009,NO_QUERY,joy_wolf,@Kwesidei not the wh
ole crew
$ head -n 5 FIT1043_Dataset
0,1467810672,Mon Apr 06 22:19:49 PDT 2009,NO_QUERY,scotthamilton,is upset that he
can't update his Facebook by texting it... and might cry as a result  School today
also. Blah!
0,1467810917,Mon Apr 06 22:19:53 PDT 2009,NO_QUERY,mattycus,@Kenichan I dived many
times for the ball. Managed to save 50% The rest go out of bounds
0,1467811184,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,ElleCTF,my whole body feels
itchy and like its on fire
0,1467811193,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,Karoli,@nationwideclass no it's
not behaving at all. i'm mad. why am i here? because I can't see you all over
there.
0,1467811372,Mon Apr 06 22:20:00 PDT 2009,NO_QUERY,joy_wolf,@Kwesidei not the whole
crew
```

Delimiter is the comma. Comma, is used to separate the column in the file .I showed the file's first five lines using the head -n 5 command in order to deduce this. Commas are used to separate the several values related to different characteristics

QUESTION 1.3

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ wc -l FIT1043_Dataset
1471793 FIT1043_Dataset
$ wc -l FIT1043_Dataset
1471793 FIT1043_Dataset
```

There are 1471793 lines in the dataset.

QUESTION A2

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ awk -F, '{print $5}' FIT1043_Dataset | sort | uniq | wc -l
626684
```

```
$ awk -F, '{print $5}' FIT1043_Dataset | sort | uniq | wc -l
626684
```

There are 626684 unique users in the dataset.

In the code, `awk -F, '{print $5}` is used to extract the fifth column from the dataset.

The unique command alone cannot work if the input is not sorted.Sort is used to sorts the output to ensure the uniq works correctly. Uniq is used to remove duplicate username, uniq only output lines that are unique in the input. wc -l is used to counts the number of unique users.

QUESTION 2.2

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ awk -F, '{print $3}' FIT1043_Dataset | head -n 1
Mon Apr 06 22:19:49 PDT 2009

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ awk -F, '{print $3}' FIT1043_Dataset | tail -n 1
Tue Jun 16 08:40:50 PDT 2009
```

```
$ awk -F, '{print $3}' FIT1043_Dataset | head -n 1
Mon Apr 06 22:19:49 PDT 2009

$ awk -F, '{print $3}' FIT1043_Dataset | tail -n 1
Tue Jun 16 08:40:50 PDT 2009
```

The date range for Twitter posts is `Mon Apr 06 22:19:49 PDT 2009 to Tue Jun 16 08:40:50 PDT 2009.`

QUESTION 2.3a

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ grep -i -w "france" FIT1043_Dataset | wc -l
1313
```

```
$ grep -i -w "france" FIT1043_Dataset | wc -l
1313
```

QUESTION 2.3b

```
$ grep -iw 'france' FIT1043_Dataset | grep -vE 'france|France' | wc -l
32
```

```
$ grep -iw 'france' FIT1043_Dataset | grep -vE 'france|France' | wc -l
32
```

32 lines not spelt exactly “france” or “France” but in other combinations of uppercase and lowercase.

grep -vE , -v will reject lines that match the pattern by inverting the match. E extended regular expressions possible, which makes more pattern matching.

'france|France'matches lines include “france” or “France” using regular expression.

QUESTION 2.2c

```
$ grep -iw "france" FIT1043_Dataset | grep -v -e "^france$" -e "^France$" >
myText.txt
```

```
0,1468015014,Mon Apr 06 23:16:04 PDT 2009,NO_QUERY,idrisjoel,is stucked in Paris and can't even travel into France
(for work)
0,1468088351,Mon Apr 06 23:38:57 PDT 2009,NO_QUERY,CathySavels,Good morning for a very rainy France  No gardening
for me today.
0,1468474003,Tue Apr 07 01:53:28 PDT 2009,NO_QUERY,Jess_18,is sad coz alison's leaving england to france tonight
0,1468693117,Tue Apr 07 03:12:11 PDT 2009,NO_QUERY,CecileNguyen,In France Today it's raining
0,1469611157,Tue Apr 07 06:54:03 PDT 2009,NO_QUERY,claire0801,Back from France with a ridiculous amount of food
but could have bought tons more. The visit made me miss living there so much
0,1553608069,Sat Apr 18 14:43:37 PDT 2009,NO_QUERY,missourie,Going to bed now Leaving france tomorrow
0,1557342015,Sun Apr 19 03:30:58 PDT 2009,NO_QUERY,kittycat1980,back from hols wanna still be in france
0,1557537700,Sun Apr 19 04:42:39 PDT 2009,NO_QUERY,Fiorinda,is back from France
0,1563934704,Mon Apr 20 00:56:59 PDT 2009,NO_QUERY,nhoizey,@atebits I'm not asleep I'm in France!!!
0,1564505864,Mon Apr 20 03:48:44 PDT 2009,NO_QUERY,richardBarley,@ikki_oo Ha well I'm leaving my home in sunny
France to fly to England which is not so sunny. So sadly not
0,1573224354,Tue Apr 21 00:13:21 PDT 2009,NO_QUERY,jcverdie,@spinningbball better than nothing  in France there's
no movies. what about price?
0,1573390846,Tue Apr 21 00:57:20 PDT 2009,NO_QUERY,annasaccone,@NatalyaFGM i just love your new pictures from
france!  i would love to go i've never ever been!
0,1573473131,Tue Apr 21 01:20:09 PDT 2009,NO_QUERY,Paulinelovejb,Miley are in Rome Demi are in Madrid And she go
to London AND THE FRANCE ????? so sad
```

QUESTION A3

```
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "usa" FIT1043_Dataset | awk -F "," '$1 == 0' | wc -l
361

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "usa" FIT1043_Dataset |awk -F "," '$1 == 2' | wc -l
0

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "usa" FIT1043_Dataset | awk -F "," '$1 == 4' | wc -l
282

$ zgrep -i -w "usa" FIT1043_Dataset | awk -F "," '$1 == 0' | wc -l
361
$ zgrep -i -w "usa" FIT1043_Dataset |awk -F "," '$1 == 2' | wc -l
0
$ zgrep -i -w "usa" FIT1043_Dataset | awk -F "," '$1 == 4' | wc -l
282
HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 0' | wc -l
596

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 2' | wc -l
0

HP PROBOOK@DESKTOP-5QV2976 MINGW64 ~/Documents/FIT1043/A3
$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 4' | wc -l
403

$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 0' | wc -l
596
$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 2' | wc -l
0
$ zgrep -i -w "canada" FIT1043_Dataset | awk -F "," '$1 == 4' | wc -l
403
```

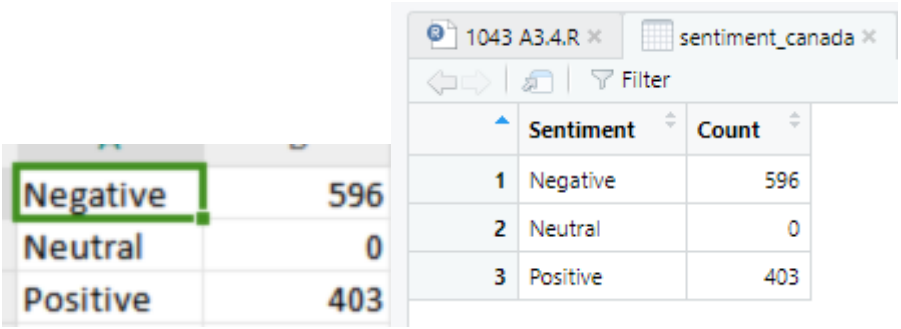
QUESTION 3.2

```
echo "Negative, <negative_count>" > sentiment-USA.csv
echo "Neutral, <neutral_count>" >> sentiment-USA.csv
echo "Postive, <postive_count>" >> sentiment-USA.csv
echo "Neutral, <neutral_count>" >> sentiment-canada.csv
echo "Negative, <negative_count>" > sentiment-canada.csv
echo "Postive, <postive_count>" >> sentiment-canada.csv
```

```
$ grep -iw "usa" FIT1043_Dataset | awk -F ',' '$1 == 0 {neg++} $1 == 2 {neu++} $1 == 4 {pos++} END {print "Negative," neg "\nNeutral," neu "\nPositive," pos}' > sentiment-USA.csv

$ grep -iw "canada" FIT1043_Dataset | awk -F ',' '$1 == 0 {neg++} $1 == 2 {neu++} $1 == 4 {pos++} END {print "Negative," neg "\nNeutral," neu "\nPositive," pos}' > sentiment-canada.csv
```

Canada



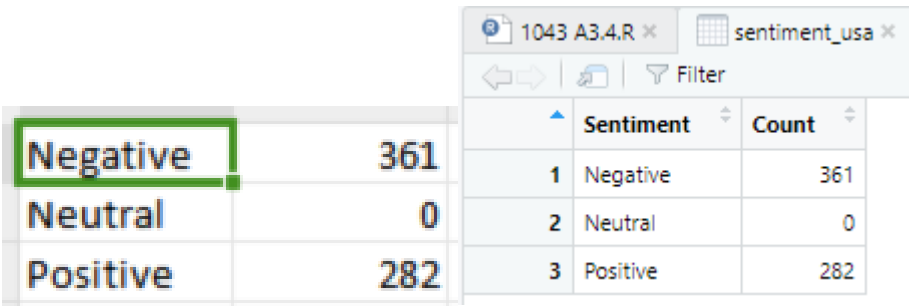
Negative	596
Neutral	0
Positive	403

1043 A3.4.R × sentiment_canada ×

Filter

	Sentiment	Count
1	Negative	596
2	Neutral	0
3	Positive	403

USA



Negative	361
Neutral	0
Positive	282

1043 A3.4.R × sentiment_usa ×

Filter

	Sentiment	Count
1	Negative	361
2	Neutral	0
3	Positive	282

QUESTION 3.3

```
# Set working directory

setwd("C:\\Users\\HP PROBOOK\\Documents\\FIT1043\\A3")

getwd()


sentiment_usa <- read.csv("sentiment-USA.csv", header = FALSE, stringsAsFactors = FALSE)

sentiment_canada <- read.csv("sentiment-canada.csv", header = FALSE, stringsAsFactors = FALSE)
```

QUESTION 3.4

```
# Manually set the column names to "Sentiment" and "Count"

colnames(sentiment_usa) <- c("Sentiment", "Count")

colnames(sentiment_canada) <- c("Sentiment", "Count")
```

Print the data to check if the headers and data are correct

print(sentiment_usa)

print(sentiment_canada)

```
> # Print the data to check if the headers and data are correct
> print(sentiment_usa)
  Sentiment Count
1  Negative   361
2   Neutral     0
3  Positive   282

> print(sentiment_canada)
  Sentiment Count
1  Negative   596
2   Neutral     0
3  Positive   403
```

Combine USA and Canada sentiment data into a matrix for plotting

Create a matrix for counts

counts <- rbind(sentiment_usa$Count, sentiment_canada$Count)

Check the counts

print(counts)

```
> # Check the counts to ensure it's a numeric matrix
> print(counts)
     [,1] [,2] [,3]
[1,]  361    0  282
[2,]  596    0  403
```

Plot the side-by-side bar chart

barplot(counts,

main = "Sentiment Comparison: USA vs Canada",

xlab = "Sentiment",

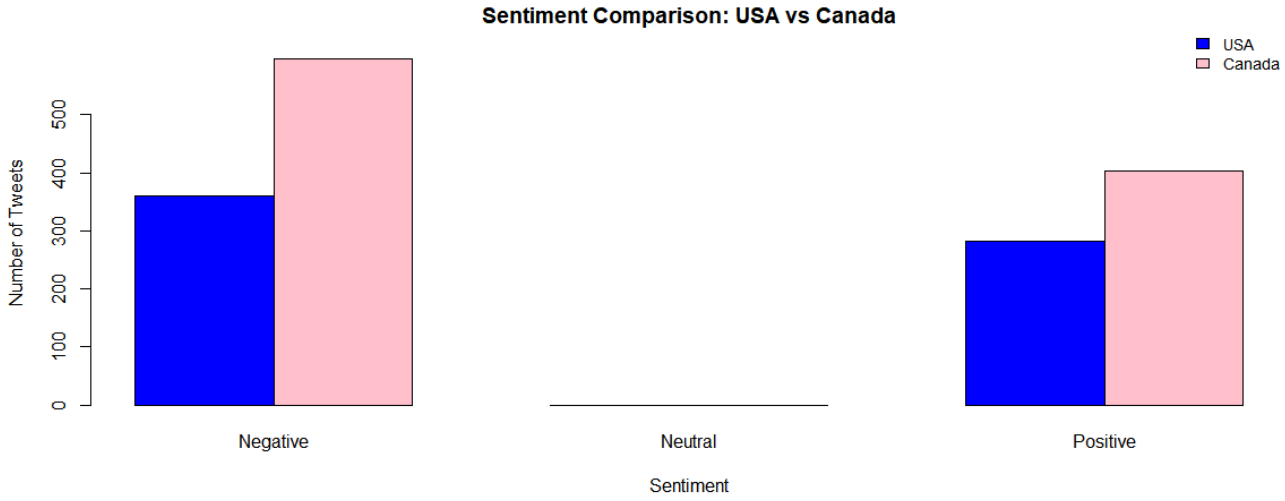
ylab = "Number of Tweets",

col = c("blue", "pink"),

names.arg = sentiment_usa$Sentiment,

beside = TRUE)

```
# Adjust the legend positioning and add a better inset for appearance
legend("topright",
       inset = c(-0.1, -0.25), # Adjust the inset to move the legend further outside
       legend = c("USA", "Canada"),
       fill = c("blue", "pink"),
       xpd = TRUE,          # Allow the legend to be drawn outside the plot area
       bty = "n",           # Remove the border around the legend
       cex = 0.9)           # Adjust the size of the legend text
```



QUESTION 3.5

Both the United States and Canada have higher negative sentiment than positive sentiment, with Canada showing higher proportion of negative sentiment (596 vs. 403) compared to the USA's ratio (361 vs. 282). Canada has more data or interaction, which may be important for comprehending local differences in emotion expression. It also means that, compared to the USA, feelings are more common and generally more negative in Canada.

QUESTION 4.1

```
grep -iw 'australia' FIT1043_Dataset | awk -F, '{print $3}' > aus_time.txt
```

```
Mon Apr 06 23:49:29 PDT 2009
Mon Apr 06 23:55:14 PDT 2009
Tue Apr 07 01:16:43 PDT 2009
Tue Apr 07 03:06:17 PDT 2009
Tue Apr 07 03:21:47 PDT 2009
Tue Apr 07 03:58:25 PDT 2009
Tue Apr 07 05:40:06 PDT 2009
Tue Apr 07 06:04:24 PDT 2009
Tue Apr 07 06:50:24 PDT 2009
Tue Apr 07 07:14:51 PDT 2009
Tue Apr 07 07:19:08 PDT 2009
Tue Apr 07 07:53:09 PDT 2009
Sat Apr 18 07:20:37 PDT 2009
Sat Apr 18 07:54:42 PDT 2009
Sat Apr 18 21:40:46 PDT 2009
Sat Apr 18 22:55:29 PDT 2009
Sat Apr 18 23:00:47 PDT 2009
Sun Apr 19 00:12:39 PDT 2009
Sun Apr 19 01:51:25 PDT 2009
Sun Apr 19 06:02:23 PDT 2009
```

QUESTION 4.2

Load the data from the text file

aus_time <- read.table("aus_time.txt", header = FALSE, stringsAsFactors = FALSE)

Combine the relevant columns into a single datetime string (ignore V5, "PDT")

AUS_time <- paste(aus_time$V1, aus_time$V2, aus_time$V3, aus_time$V4, aus_time$V6)

Remove "PDT" using gsub

remove_PDT <- gsub("PDT", "", AUS_time)

Trim any extra spaces that might appear

remove_PDT <- gsub("\\s+", " ", remove_PDT)

Convert the cleaned datetime strings to POSIXlt format

#convert from string value using strptime () function

aus_times_converted <- strptime(remove_PDT, format = "%a %b %d %H:%M:%S %Y")

#%a: Abbreviated weekday name

#%b: Abbreviated month name

#%d: Day of the month as decimal number

#%H: Hours as decimal number

#%M: Minute as decimal number

#%S: Second as integer

Add the converted datetime as a new column

aus_time$Converted_Time <- aus_times_converted

Extract only the date part from the datetime

aus_time$Date <- as.Date(aus_time$Converted_Time)

1043 A3.4.R* | sentiment_canada | aus_time | aus_time_parsed

Filter

	V1	V2	V3	V4	V5	V6	Converted_Time	Date
1	Mon	Apr	6	23:49:29	PDT	2009	*2009-04-06 23:49:29*	2009-04-06
2	Mon	Apr	6	23:55:14	PDT	2009	*2009-04-06 23:55:14*	2009-04-06
3	Tue	Apr	7	01:16:43	PDT	2009	*2009-04-07 01:16:43*	2009-04-07
4	Tue	Apr	7	03:06:17	PDT	2009	*2009-04-07 03:06:17*	2009-04-07
5	Tue	Apr	7	03:21:47	PDT	2009	*2009-04-07 03:21:47*	2009-04-07
6	Tue	Apr	7	03:58:25	PDT	2009	*2009-04-07 03:58:25*	2009-04-07
7	Tue	Apr	7	05:40:06	PDT	2009	*2009-04-07 05:40:06*	2009-04-07
8	Tue	Apr	7	06:04:24	PDT	2009	*2009-04-07 06:04:24*	2009-04-07
9	Tue	Apr	7	06:50:24	PDT	2009	*2009-04-07 06:50:24*	2009-04-07
10	Tue	Apr	7	07:14:51	PDT	2009	*2009-04-07 07:14:51*	2009-04-07

QUESTION 4.3

```
# Count the number of tweets for each day
tweet_counts_by_day <- table(aus_time$Date)


# Create the histogram without x-axis label (xlab left blank)
barplot(tweet_counts_df$Freq,
        names.arg = tweet_counts_df$Var1,
        main = "Number of Tweets per Day",
        xlab = "",  # I add it manually with mtext
        ylab = "Number of Tweets",
        col = "lightblue",
        las = 2,         # Make x-axis labels vertical
        cex.names = 0.8,  # Adjust size of x-axis labels
        border = "black") # Add black border around bars to improve visibility


# Manually add the x-axis label with more space at the bottom
mtext("Date", side = 1, line = 5, cex = 1.1)
```

1043 A3.4.R* × tweet_counts_df ×

Filter

	Var1	Freq
1	2009-04-06	3
2	2009-04-07	20
3	2009-04-17	2
4	2009-04-18	15
5	2009-04-19	16
6	2009-04-20	23
7	2009-04-21	16
8	2009-05-01	10
9	2009-05-02	25
10	2009-05-03	34
11	2009-05-04	23
12	2009-05-09	7
13	2009-05-10	18
14	2009-05-11	4
15	2009-05-13	5
16	2009-05-14	13
17	2009-05-16	13

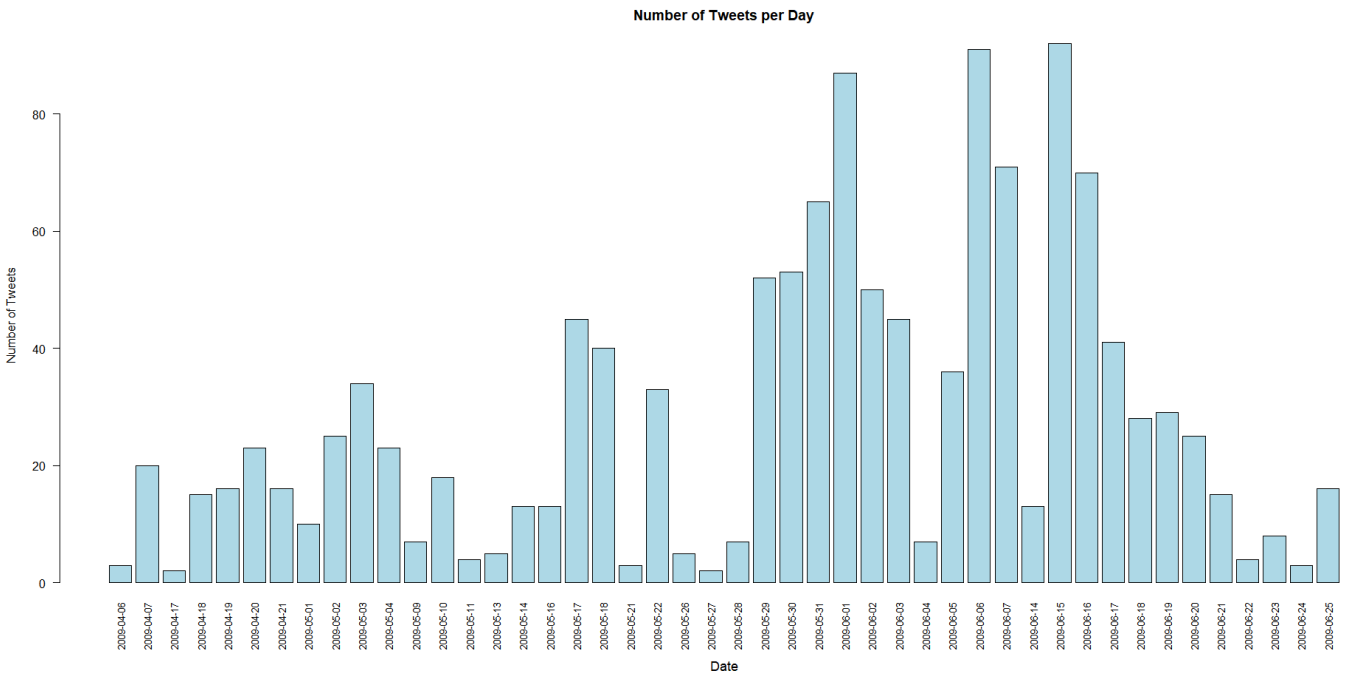
Showing 1 to 17 of 46 entries, 2 total columns

R | Global Environment

Data

aus_time	1283 obs. of 8 variables
aus_time_parsed	POSIXlt[1:1759], format: NA NA NA NA NA NA NA NA ...
aus_times_converted	POSIXlt[1:1283], format: "2009-04-06 23:49:29" ...
combined_data	int [1:2, 1:3] 2573 698 NA NA 3112 541
counts	int [1:2, 1:3] 361 596 0 0 282 403
sentiment_canada	3 obs. of 2 variables
sentiment_usa	3 obs. of 2 variables
timestamps	POSIXlt[1:1759], format: "2009-04-06 23:49:29" "2009-04-06 23:55:14" "2009-04-07 01:16:43" ...
tweet_counts_df	46 obs. of 2 variables

Values



Distribution of Tweets

```
> print(summary(tweet_counts_df$Freq))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    7.25   19.00   27.89   40.75   92.00
```

Min 2.00

The minimum number of tweets on any day is 2

1st Quartile – 7.25

25% of the days have a tweet count of 7.25

Median -19

The median number of tweets per day is 19

Mean 27.89

Average 17.89 tweets per day

3RD quartile – 40.75

75% of the days have 40.75 tweets

Max – 92

The highest number of tweets on single day is 92

These statistics reveal that the distribution of tweet counts is right-skewed, as the mean (27.89) is greater than the median (19). This suggests that a few days with high tweet counts are pulling the average upward.

The histogram shows a significant variation in the number of tweets throughout the examined time, with some days showing little activity (less than 20 tweets) and others exhibiting sudden spikes, surpassing 80 tweets. Late May and early June have the most activity, especially on June 1, 6, and 15, which may indicate that significant events or trends were the cause of these spikes. On the other hand, April exhibits comparatively low activity, with daily counts between 10 and 30 tweets, and late June sees a similar decline in interaction. Between May 10 and May 19, there is also a noticeable decrease in activity, with tweet counts dropping below 20. The distribution of the data is mostly right skewed, with occasional spikes that are probably related to outside campaigns or events. According to the third quartile figure of 40.75, fewer than 40 tweets were sent on 75% of the days, with the higher end of the distribution being driven by a small number of outliers, such as a maximum count of 92.