Name: Abrar Fauzan Hamzah
ID: 28551494

# Assignment 3

## Part A:

**1.** 2.1GB

```
➢ ls -lh Twitter_Data_1.txt
```

**2.** **delimiter** : TAB & **columns** : 3

```
➢ head -1 Twitter_Data_1.txt | less

#and to find the delimiter I use
➢ "<ctrl+V>+<TAB>"
```

**3.** From using the code above (Q2) we could see that the other column names are Username, Date/Time, Tweet

**4.** There are 15089920 lines.

```
➢ wc -l Twitter_Data_1.txt
```

**5.** The range of that is from February 11 2014 to February 18 2014

```
➢ cut -f 3 Twitter_Data_1.txt
```

**6.** 8977904 unique usernames

```
➢ cut -f 2 Twitter_Data_1.txt | sort |
uniq -c | wc -l
```

Name: Abrar Fauzan Hamzah
ID: 28551494

**7.** **Tweet** : "RT @aedan_smith: Be interesting to see the detail on this one: BBC News - Donald Trump loses offshore wind farm challenge http://t.co/qAcG… "

```
➢ cut -f 4 Twitter_Data_1.txt | grep "Donald Trump"
```

**Date/Time** : Tuesday, 11 February 2014 12:28:36
To get the date I use grep with the tweet that I have found.

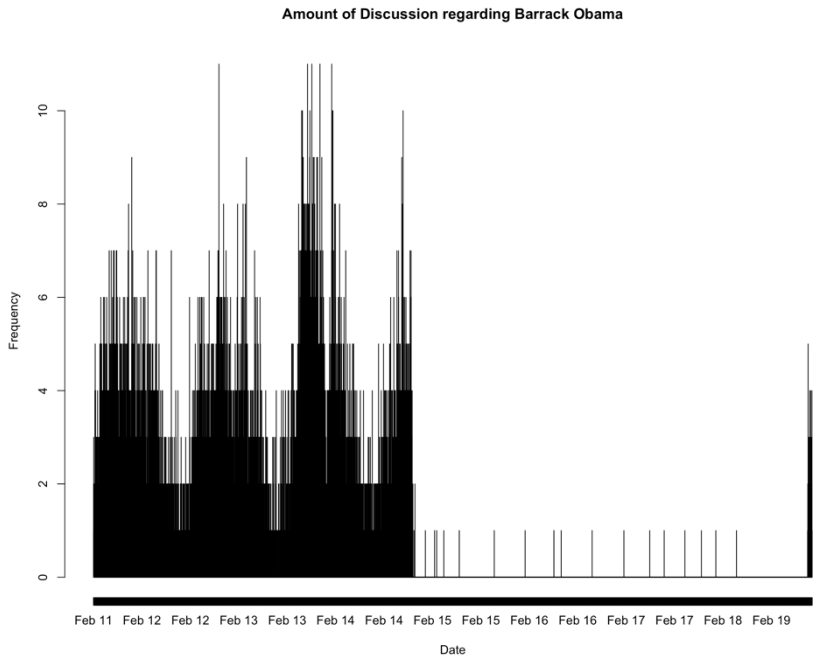
**8.** **No**, because there might be misspelled words like "Donld Trmp" or there might be someone who changed the characters to symbols like "D0n4ld Trump", or even there might be a spelling of "Donald Trump" in other languages. Also, we only search with capital D and T, we did not search when all characters are lower case or upper case.

## Part B:

1. 10909 Tweets with "Obama".
   - ➢ $ grep "Obama" Twitter_Data_1.txt | wc -l

2. Format = "%a %b %d %H:%M:%S %z %Y"

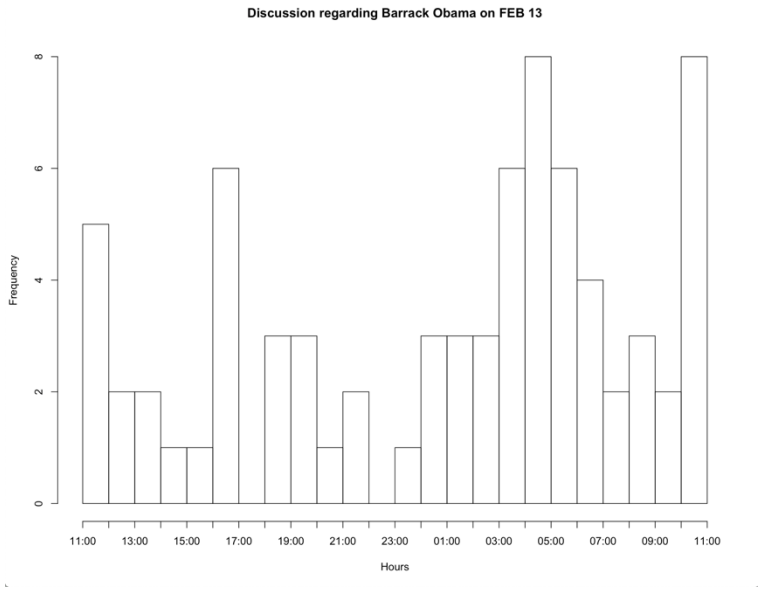
3. After using the strptime() function, we plot a histogram with bins = 10000.



Name: Abrar Fauzan Hamzah
ID: 28551494

**4.** From the histogram above, we can see that before February 15, there are a lot of tweets regarding Barrack Obama also a pattern of up and down each day. However , after February 15, it significantly dropped, almost no one is discussing about him.

**5.**




From this histogram we can see that the time that most people were discussing Barrack Obama is around 3am to 7am in the morning of Thursday, February the 13th.

Name: Abrar Fauzan Hamzah
ID: 28551494

## Part C:

For this task, we will evaluate the relation between Rings and Diameter of an Abalone to determine the sex of it. There are 3 sex of abalone which are Male, Female and Infant. An infant is classified as a sex because it is hard to determine the sex of abalone when it is still at young age. The method that we use will be k-mean clustering with k =3 since there are 3 sexes.

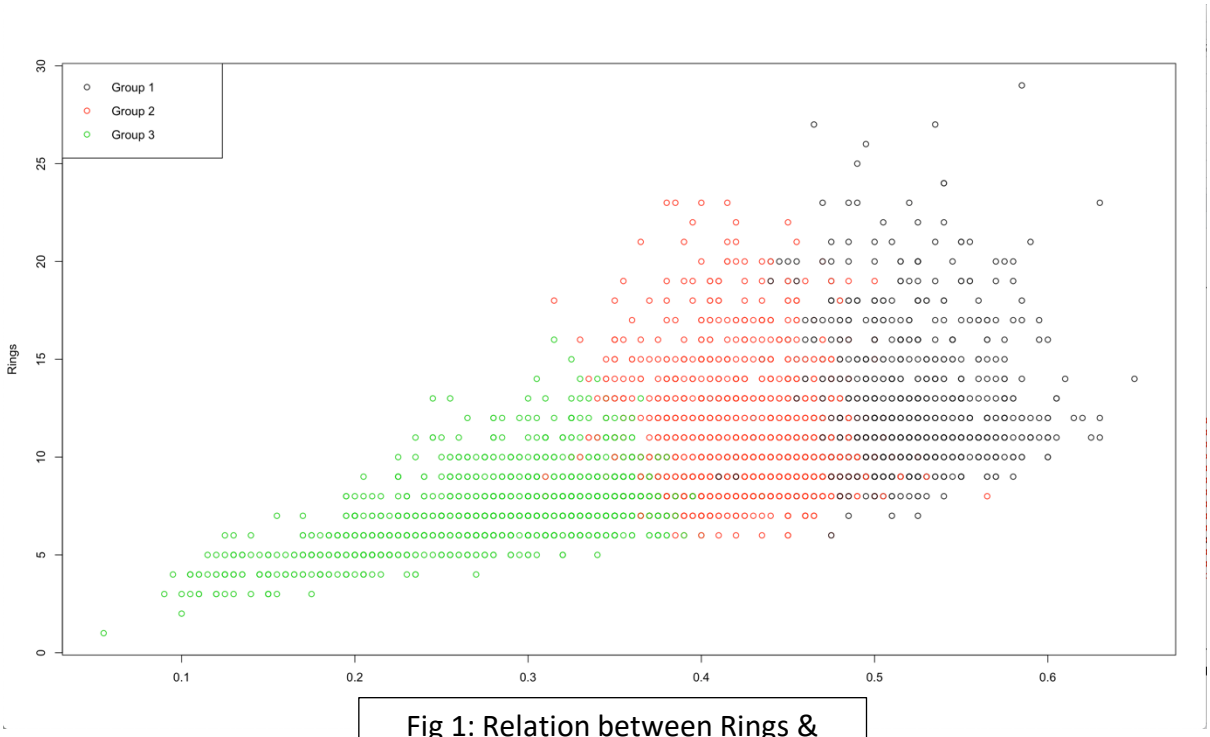



Fig 1: Relation between Rings & Diameter using k-means clustering with k=3

From figure 1, we can see that group 3 (green) is grouped as Infant, group 2 as male and 3 as female. The graph shows us that the more rings and the bigger the diameter of an abalone it shows that it is a adult male. While an abalone with Rings ~below 15 with a diameter under 0.4, an abalone is categorised as an infant.

Name: Abrar Fauzan Hamzah
ID: 28551494

Now, lets see the graph with the original data without using k-means clustering to compare the accuracy of our k-mean clustering method.

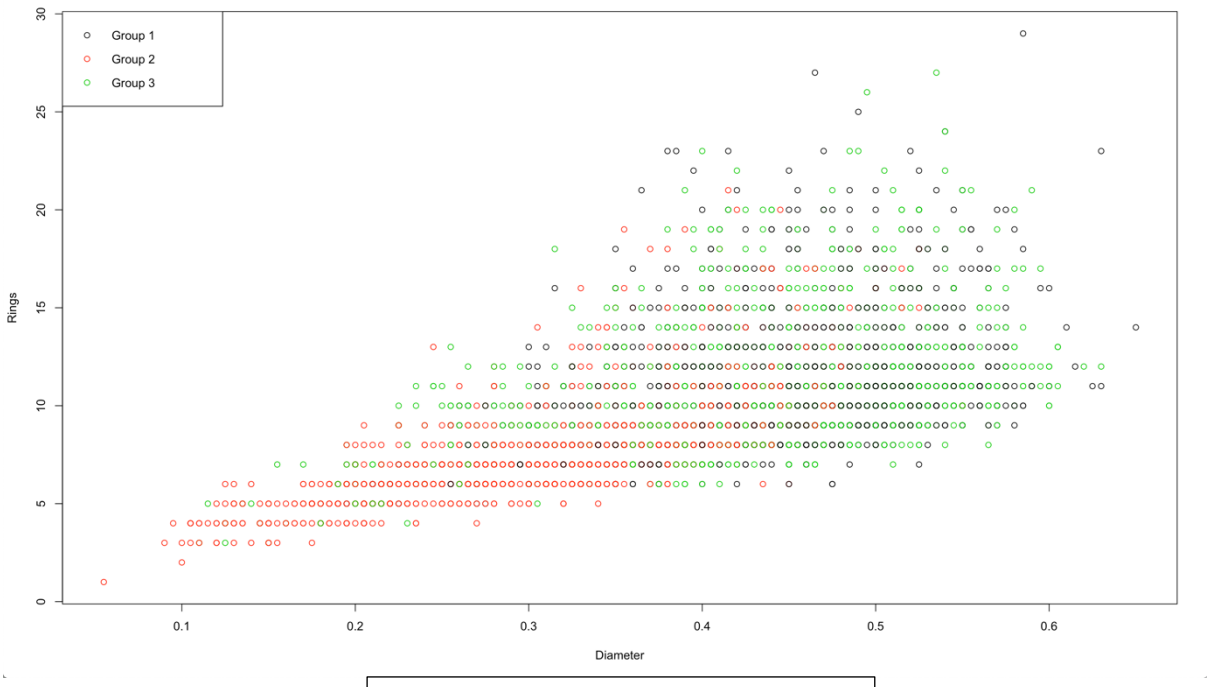



Fig 2: Relation between Rings & Diameter using the original data

In figure 2, we see that Group 1 & 3, which are Male and Female, are similar. The data that we have indicates that there is no significant relation between Rings and Diameter to be able to determine the sex of an adult Abalone.

In conclusion, it is best to use figure 1 to determine between adult and infant since the difference between figure 1 and 2 is small.

Name: Abrar Fauzan Hamzah
ID: 28551494

## R code for PartB

```
#Q3 Part B:

#reading the date file that contains "Obama"
df <- read.csv("/Users/abrarfauzanhamzah/Desktop/yes.txt", header = FALSE)
View(df)
names(df)<- "Date"
"Convert string of dates to correct class"
res <- strptime(df[,1], format="%a %b %d %H:%M:%S %z %Y")


#plot histogram
hist(res, breaks=10000, freq=TRUE, xlab = "Date", main = "Amount of Discus
sion regarding Barrack Obama")


#Q4 Part B:

#read dates that has most discussion on Obama
feb13 <- read.csv("/Users/abrarfauzanhamzah/Desktop/Feb13.txt", header = F
ALSE)

names(feb13)<- "Date"
View(feb13)
feb <- strptime(feb13[,1], format="%a %b %d %H:%M:%S %z %Y")

hist(feb, breaks='hours', freq=TRUE, xlab = "Hours", main = "Discussion re
garding Barrack Obama on FEB 13")
```