FIT 1043 Assignment 3

Monash ID: 34076492

Name: Daffa Fariq Prasetyo

Task A1: inspecting the data.

1) Copying the file into the cygin.

$ cp /cygdrive/c/Users/Daffa/Downloads/corona_tweets.csv.gz ~/

This was done to move the current directory in the right file of where the corona_tweets.csv is located at.

File size of the compressed csv.

$ ls -lh corona_tweets.csv.gz

'ls' command that is used to list the directory contents.

'-l' tells 'ls' to use long listing format which provides detailed information about each file.

'-h' modifies the file size information to be displayed in a more readable format.

The File size is 118MB while being compressed.

```
daffa@laptop ~
$ ls -lh corona_tweets.csv.gz
-rwx------ 1 daffa daffa 118M May 23 15:44 corona_tweets.csv.gz
```

2) Header names separated by "tab."

This code is to find the headers that are already separated by tab.

$ zcat corona_tweets.csv.gz | head -n 1

'zcat' command that is used to decompress and display the contents of the gzip-compressed file without extracting it to the disc.

'head -n 1' command that is used to read the first 'n' lines from its input and outputs them. The '-n 1' specifies that only the first line should be the output.

Header names are:

Created, Tweet_ID, Text, User_ID, User, User_Location, Followers_Count, Friends_Count, Geo, Place_Type, Place_Name, Place_Country, Language.

```
daffa@laptop ~
$ zcat corona_tweets.csv.gz | head -n 1
Created Tweet_ID        Text    User_ID User    User_Location   Followers_Count Friends_Count   Geo     Place_Type      Place_Name      Place_Country   Language
```

3) Number of lines.

$ zcat corona_tweets.csv.gz | wc -l

'zcat' command that is used to decompress and display the contents of the gzip-compressed file without extracting it to the disc.

'wc-l' stands for word count which can also be used to count the line, words and characters. '-l' makes it so that it counts the number of lines in the decompressed data.

```
daffa@laptop ~
$ zcat corona_tweets.csv.gz | wc -l
1143559
```

There are 1,143,559 lines of code.

Task A2: Investigating the information.

1) Unique twitter usernames.

$ zcat corona_tweets.csv.gz | awk -F '\t' '{print $4}' | sort | uniq | wc -l
Using the uniq removes the adjacent duplicate lines, so data must be sorted first.

'zcat': command that is used to decompress and display the contents of the gzip-compressed file without extracting it to the disc.

awk -F '\t' '{print $4}': sets the field separated to tab '\' and prints the 4th field of each line.

sort | uniq: sorts the username alphabetically and uniq removes the adjacent duplicate lines.

'wc -l': counts the number of lines that are unique.

There are 641,976 unique usernames in the twitter file.

```
daffa@laptop ~
$ zcat corona_tweets.csv.gz | awk -F '\t' '{print $4}' | sort | uniq | wc -l
641976
```

2) sub-questions

A) ***Tweets mentioning the word "vaccine" in any combination.***

zcat corona_tweets.csv.gz | grep -i 'vaccine' | wc -l

'zcat': command that is used to decompress and display the contents of the gzip-compressed file without extracting it to the disc.

'grep -i 'vaccine'': Search for the line with the word "vaccine" , '-l' makes the search case insensitive.

'wc -l': counts the number of lines that are unique.
There are 19,483 tweets mentioning 'vaccine'.

```
daffa@laptop ~
$ zcat corona_tweets.csv.gz | grep -i 'vaccine' | wc -l
19483
```

B) ***Not spelt exactly "vaccine" or "Vaccine"***

$ zcat corona_tweets.csv.gz | grep -i -E '\bv[^]*a[^]*c[^]*c[^]*i[^]*n[^]*e\b' | wc -l

grep -i -E '\bv[^]*a[^]*c[^]*c[^]*i[^]*n[^]*e\b'

'I': Searches for the word in a case insensitive way

'-E': Enables extended regular expressions.

'\bv[^]*a[^]*c[^]*c[^]*i[^]*n[^]*e\b': Searches the expression from the boundary 'b' and '\b' that matches the phrase of any combination from the upper and lower case from the characters, "accine"

```
daffa@laptop ~
$ zcat corona_tweets.csv.gz | grep -i -E '\bv[^ ]*a[^ ]*c[^ ]*c[^ ]*i[^ ]*n[^ ]*
e\b' | wc -l
16773
```

Out of the 19,483 tweets, 16,773 of them are not in the form "vaccine" or "Vaccine".

C) ***Output the lines into a file called "Result.txt"***

$ zcat corona_tweets.csv.gz | grep -i -E '\bv[^]*a[^]*c[^]*c[^]*i[^]*n[^]*e\b' > Result.txt

> Result.txt

redirects the output of the grep into Results.txt into the home page, if I wanted it in my downloads:

Task A3:

1)Format = zcat corona_tweets.csv.gz | awk -F '\t' '$7 condition {print $4}' | sort | uniq | wc -l

'zcat': handle gzip compressed files.

'awk -F '\t' '$7 condition {print $4}': filters the data based on the 7th field and prints out the twitter user ID (4th field).

'sort | uniq': sorts the username alphabetically and uniq removes the adjacent duplicate lines.

'wc -l': counts the number of lines that are unique.

a) Less than or equal to 1500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 <= 1500 {print $4}' | sort | uniq | wc -l
498,480 users

b) 1501 to 2500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 1501 && $7 <= 2500 {print $4}' | sort | uniq | wc -l
43,891 users

c) 2501 to 3500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 2501 && $7 <= 3500 {print $4}' | sort | uniq | wc -l
23,620 users

d) 3501 to 4500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 3501 && $7 <= 4500 {print $4}' | sort | uniq | wc -l
15,165 users

e) 4501 to 5500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 4501 && $7 <= 5500 {print $4}' | sort | uniq | wc -l
9,297 users

f) 5501 to 6500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 5501 && $7 <= 6500 {print $4}' | sort | uniq | wc -l
6,848 users

g) 6501 to 7500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 6501 && $7 <= 7500 {print $4}' | sort | uniq | wc -l
5076 users

h) 7501 to 8500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 7501 && $7 <= 8500 {print $4}' | sort | uniq | wc -l
3,855 users

i) 8501 to 9500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 >= 8501 && $7 <= 9500 {print $4}' | sort | uniq | wc -l
3,072 users

j) More than 9500.

$ zcat corona_tweets.csv.gz | awk -F '\t' '$7 > 9500 {print $4}' | sort | uniq | wc -l
32,772 users

2) CSV file

```
Range,Frequency
<=1500,498480
1501 to 2500,47176
2501 to 3500,23620
3501 to 4500,15165
4501 to 5500,9297
5501 to 6500,6848
6501 to 7500,5076
7501 to 8500,3855
8501 to 9500,3072
> 9500,32772
```

Or from the cygin terminal, (not sure what the question meant)

```
echo "Follower Range,Number of Twitter Users" > follower_ranges.csv
echo "<= 1500,$less_than_1500" >> follower_ranges.csv
echo "1501 to 2500,$range_1501_2500" >> follower_ranges.csv
echo "2501 to 3500,$range_2501_3500" >> follower_ranges.csv
echo "3501 to 4500,$range_3501_4500" >> follower_ranges.csv
echo "4501 to 5500,$range_4501_5500" >> follower_ranges.csv
echo "5501 to 6500,$range_5501_6500" >> follower_ranges.csv
echo "6501 to 7500,$range_6501_7500" >> follower_ranges.csv
echo "7501 to 8500,$range_7501_8500" >> follower_ranges.csv
echo "8501 to 9500,$range_8501_9500" >> follower_ranges.csv
echo "> 9500,$more_than_9500" >> follower_ranges.csv
```

3) Reading the file

```
data <- read.csv("C:/Users/daffa/Downloads/follower_ranges.csv")
```

4) bar chart

#Making the bar chart using ggplot

library(ggplot2)

#X order is incorrect somehow so I must do this.

data$Follower.Range <- factor(data$Follower.Range, levels = c("<= 1500", "1501 to 2500", "2501 to 3500", "3501 to 4500", "4501 to 5500", "5501 to 6500", "6501 to 7500", "7501 to 8500", "8501 to 9500", "> 9500"))

#Plotting it into a bar chart

plot <- ggplot(data, aes(x = Follower.Range, y = Number.of.Twitter.Users)) +

geom_bar(stat = "identity", fill = "lightblue") +

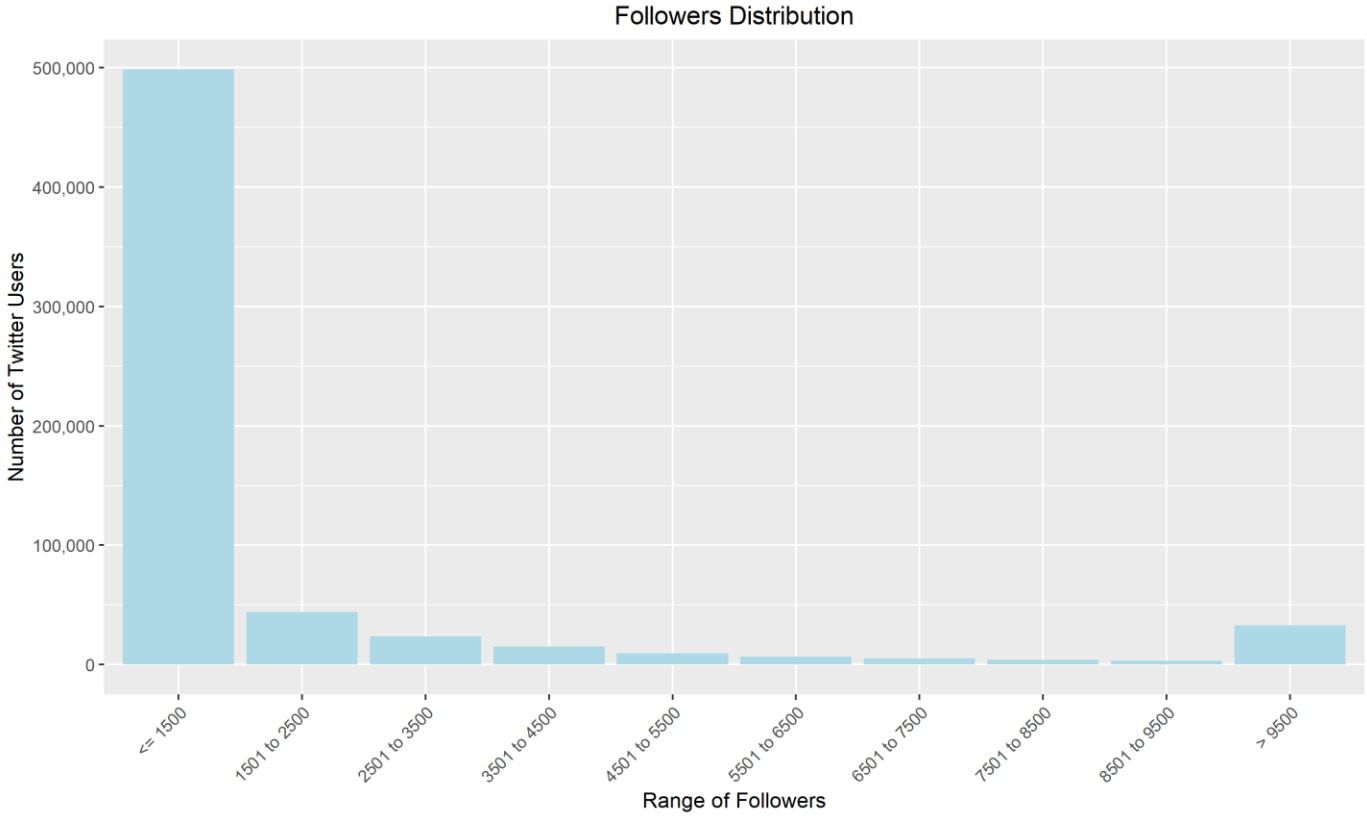
labs(title = "Followers Distribution", x = "Range of Followers", y = "Number of Twitter Users") +

theme(axis.text.x = element_text(angle = 45)) +

scale_y_continuous(labels = scales::comma)

#Making it into a PNG

ggsave("C:/Users/daffa/Downloads/follower_counts.png", plot, width = 10, height = 6)



Task A4:

1) Tweets with no “RT @”.

zcat corona_tweets.csv.gz | grep -v "RT @" | gzip > non_retweet_tweets.csv.gz

- ‘grep -v "RT @"’ – search for lines that do not contain the pattern "RT @". ‘-v’ inverts the match.
- ‘gzip > non_retweet_tweets.csv.gz’ – compress the filtered data and redirects the output of the command to the file ‘non_retweet_tweets.csv.gz’.

2) Same process as A3.1 to A3.2

Less than or equal to 1500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 <= 1500 {print $4}' | sort | uniq | wc -l

156,973 users

1501 to 2500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 1500 && $7 <= 2500 {print $4}' | sort | uniq | wc -l

16,060 users

2501 to 3500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 2500 && $7 <= 3500 {print $4}' | sort | uniq | wc -l

9,013 users

3501 to 4500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 3500 && $7 <= 4500 {print $4}' | sort | uniq | wc -l

6,069 users

4501 to 5500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 4500 && $7 <= 5500 {print $4}' | sort | uniq | wc -l

3,870 users

5501 to 6500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 5500 && $7 <= 6500 {print $4}' | sort | uniq | wc -l

2,965 users

6501 to 7500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 6500 && $7 <= 7500 {print $4}' | sort | uniq | wc -l

2,186 users

7501 to 8500

zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 7500 && $7 <= 8500 {print $4}' | sort | uniq | wc -l

1,726 users

8501 to 9500

```
zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 8500 && $7 <= 9500 {print $4}' | sort | uniq | wc -l
```

1,425 users

More than 9500

```
zcat non_retweet_tweets.csv.gz | awk -F '\t' '$7 > 9500 {print $4}' | sort | uniq | wc -l
```

17,624

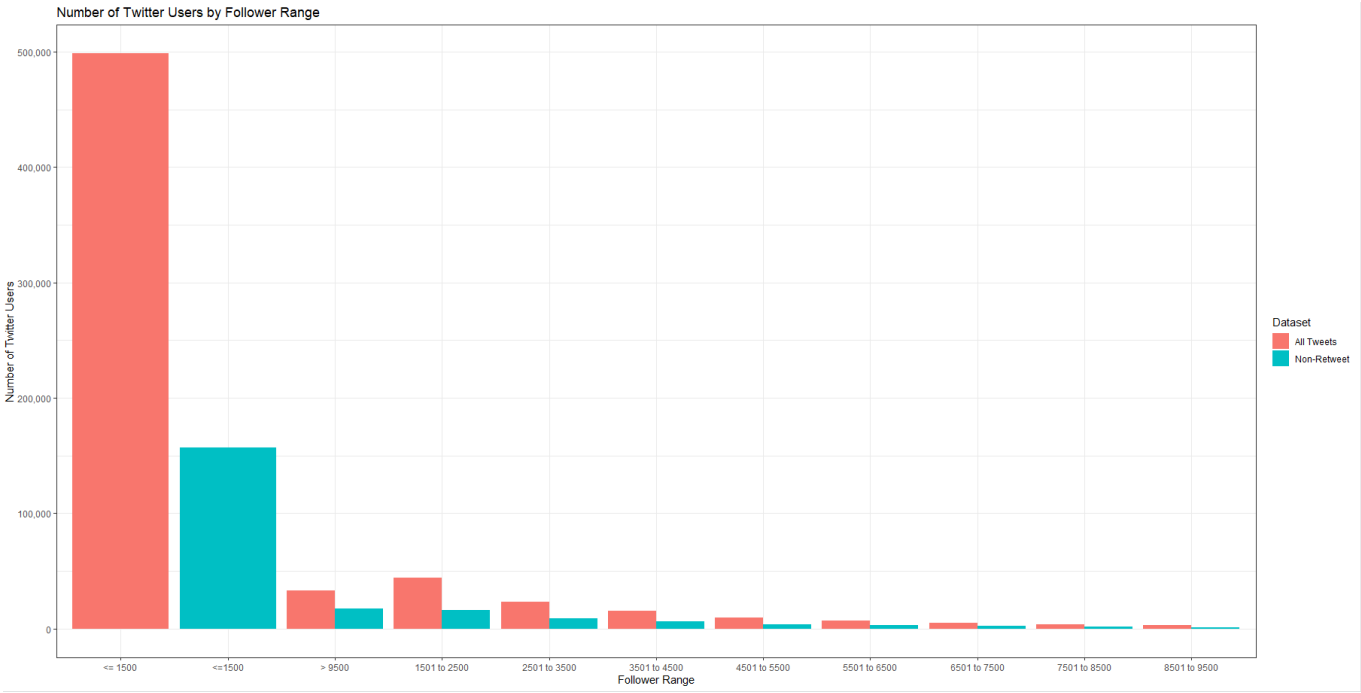
3 and 4) Rcode for a side-by-side bar graph

```
library(ggplot2)
Non_RT_data <- read.csv("C:/Users/daffa/Downloads/non_retweet_tweets.csv")
data <- read.csv("C:/Users/daffa/Downloads/follower_ranges.csv")

#New column to identify the datasets
Non_RT_data$Dataset <- "Non-Retweet"
data$Dataset <- "All Tweets"

#Combine the datasets
combined_data <- rbind(Non_RT_data, data)

#Create the bar chart
ggplot(combined_data, aes(x = factor(`Follower.Range`), y = `Number.of.Twitter.Users`, fill = Dataset)) +
  geom_col(position = "dodge") +
  labs(title = "Number of Twitter Users by Follower Range",
       x = "Follower Range",
       y = "Number of Twitter Users") +
  theme_bw() +
  scale_y_continuous(labels = scales::comma)
```



Findings

This plotted chart presents the distribution of Twitter users mention the word vaccine, categorized by their follower range, and distinguishes them between all the tweets and the non-retweets tweets.

Majority of users have a low follower counts where most of the user's mentioning vaccine falls into the "<= 1500 followers" range. This group has the highest number of users portraying that most of the conversation around vaccines is driver by user with lower follower counts. Within this group, there are significantly more total tweets than non-retweets suggesting that retweets are a major component of the vaccine-related conversation with users with low follower counts.

There is a sharp decline in the number of users as the follower count increases. For example, the number of users in the 1501 to 2500 followers and the 2501 to 3501 followers ranges is significantly lower than those in the <= 1500 followers' range. This pattern continues consistently across all higher follower ranges, indicating that fewer users with higher follower counts are discussing vaccine or that these discussions are captured less frequently in this dataset.

Consistent pattern across non-retweet tweets where the number of conversations about vaccines are less across all follower ranges indicating that the original content about vaccines is proportionally distributed across different follower ranges similarly to overall activity.

In the higher follower ranges, e.g., 6501 to 7500 and above, both total retweets and non-retweets are minimal showing that users with larger following are less engaged in tweeting about vaccines or scared to mention anything about vaccines as they will lose their followers and scared, they will be taken of the platform for misinformation maintaining their reputation.

This whole finding shows that vaccine related discussion on twitter is predominantly driver by users with smaller following with retweets playing a crucial role in amplifying their message.