RProject_group(Leysa,Calambro)

Camilo Leysa

2024-12-05

Loading necessary libraries

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.4.2
```

```r
library(textdata)
```

```
## Warning: package 'textdata' was built under R version 4.4.2
```

```r
library(ggplot2)
library(dplyr)
library(lubridate)
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.4.2
```

```
## Loading required package: RColorBrewer
```

```r
library(RColorBrewer)
```

Loading the data

```r
the_data <- read.csv("tweetsDF.csv")
```

Removing something from the file that preventing the file to knit.

```r
the_data <- the_data %>%
  mutate(across(where(is.character), ~ gsub(" ", "", .)))
```

Preview of the data

```r
head(the_data)
```

```
##   X    screenName
## 1 1      whourj31
## 2 2       nnainot
## 3 3   febry_sri_M
## 4 4 telehuntwatch
## 5 5    Typing0824
## 6 6   niccijsmith
##
## 1          A soldier angry at the support fund consolation money for the bereaved family of the Itae
## 2                                                                                       Nah this Ita
## 3
## 4 TRANSLATION :\nSeoul residents lay flowers at a makeshift memorial near the site of the crush in I
## 5  The Itaewon stampede incident really caught me off guard. Makes me notice how important it is to
## 6  "What to do about my child? What to do about my child?" Park Ga-young’s mother, Choi Seon-mi, sai
##               created
## 1 2022-10-30 23:59:43
## 2 2022-10-30 23:59:32
## 3 2022-10-30 23:59:31
## 4 2022-10-30 23:59:28
## 5 2022-10-30 23:59:20
## 6 2022-10-30 23:59:04
##                                                                            statusSource
## 1                <a href="https://www.fs-poster.com/" rel="nofollow">FS_Poster_App</a>
## 2 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 3 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 4                            <a href="https://ruprop.live" rel="nofollow">telehunt</a>
## 5 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 6   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
##      Created_At_Round tweetSource
## 1 2022-10-31 00:00:00      others
## 2 2022-10-31 00:00:00     android
## 3 2022-10-31 00:00:00     android
## 4 2022-10-31 00:00:00      others
## 5 2022-10-31 00:00:00     android
## 6 2022-10-31 00:00:00      iphone
```

```r
tail(the_data)
```

```
##           X    screenName
## 58081 58081  lovemaselfbb
## 58082 58082 seokjlnswhore
## 58083 58083 bjorktheworld
## 58084 58084    and_katara
## 58085 58085    s_meraldo7
## 58086 58086     N_ick_y28
##
## 58081                                                                       @sbmptnfess Start-gaho (
## 58082                                                   itaewon class was so overhyped for no reason
## 58083                                                                                               @
## 58084
## 58085                                                                   @philtaetos Wait y’all don’t
## 58086 @philtaetos Ok I'm gonna break my silence.\nI don't like itaewon class somehow I started it I
##                   created
## 58081 2022-10-28 00:27:14
## 58082 2022-10-28 00:16:17
## 58083 2022-10-28 00:15:24
## 58084 2022-10-28 00:09:23
## 58085 2022-10-28 00:03:08
## 58086 2022-10-28 00:01:06
##                                                                               statusSource
## 58081 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 58082 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 58083   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 58084 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 58085   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 58086 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
##          Created_At_Round tweetSource
## 58081 2022-10-28 00:00:00     android
## 58082 2022-10-28 00:00:00     android
## 58083 2022-10-28 00:00:00      iphone
## 58084 2022-10-28 00:00:00     android
## 58085 2022-10-28 00:00:00      iphone
## 58086 2022-10-28 00:00:00     android
```

Getting the total number of each TweetSource and saving as Data Frame

```
tweet_count_df <- as.data.frame(table(the_data$tweetSource)) %>%
  rename(Source = Var1, Count = Freq)
tweet_count_df
```

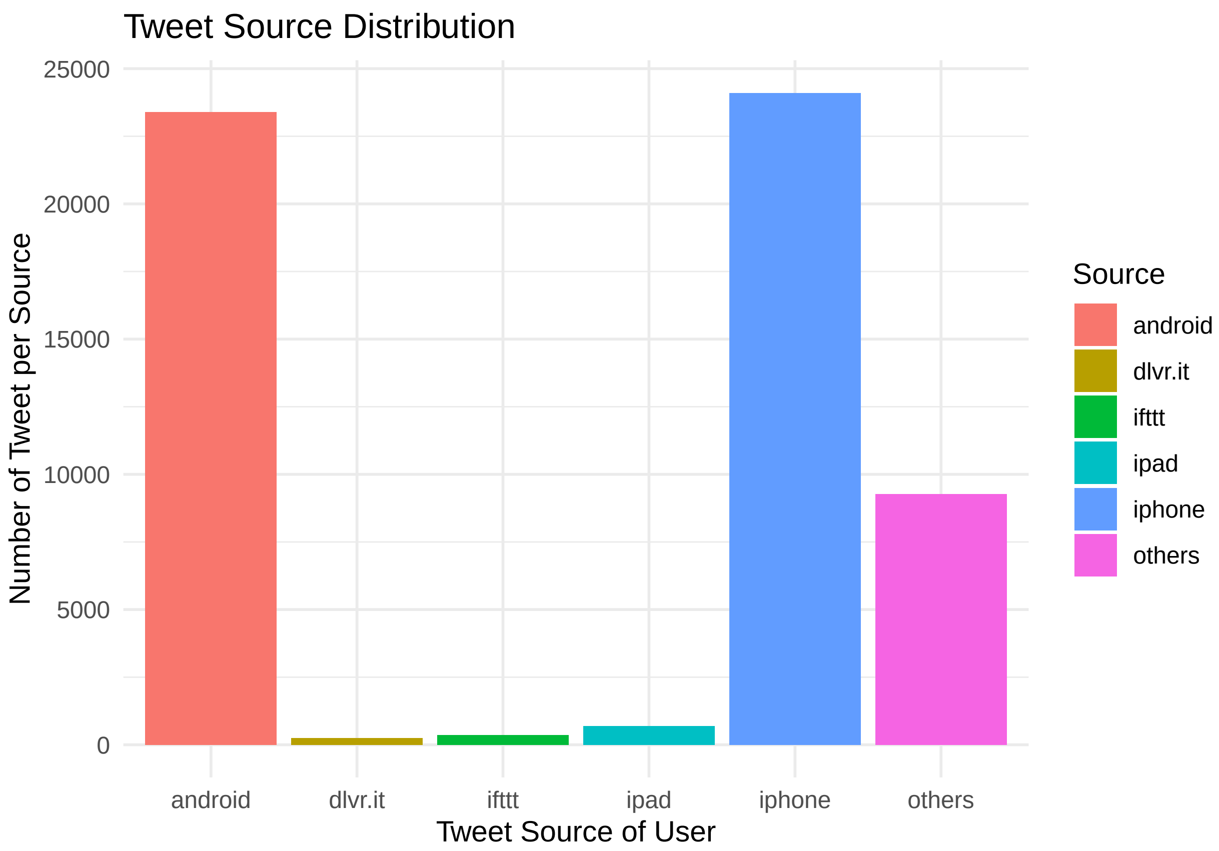
```
##    Source Count
## 1 android 23399
## 2 dlvr.it   241
## 3   ifttt   364
## 4    ipad   698
## 5  iphone 24106
## 6  others  9278
```

Saving the data as CSV

```
write.csv(tweet_count_df, file = "TweetSource.csv", row.names = FALSE)
```

Plotting the data of the Tweet_Source

```
ggplot(tweet_count_df, aes(x = Source, y = Count, fill = Source)) + geom_bar(stat = "identity") + labs(
```



```
#The plot shows that most tweets come from iPhone users, followed by Android users, and then others out
```

Getting the number of tweets using the time and plotting it

Accessing the column of time in the csv file.

```
the_data$created <- ymd_hms(the_data$created, tz = "Asia/Manila")
tweet_rtime <- as.POSIXct(the_data$created, tz = "")
head(tweet_rtime)
```

```
## [1] "2022-10-30 23:59:43 +08" "2022-10-30 23:59:32 +08"
## [3] "2022-10-30 23:59:31 +08" "2022-10-30 23:59:28 +08"
## [5] "2022-10-30 23:59:20 +08" "2022-10-30 23:59:04 +08"
```

Getting the number of tweets every 8 hours #I use 8 hours interval since the data in the dataset has a 3 day data of tweets.

```
hour_tweet <- the_data %>%
  mutate(interval = floor_date(created, "8 hours")) %>%
  group_by(interval) %>%
  summarise(count = n(), .groups = 'drop')
  sum(is.na(the_data$created))
```

```
## [1] 0
```

```
hour_tweet
```

```
## # A tibble: 9 x 2
##   interval            count
##   <dttm>              <int>
## 1 2022-10-28 00:00:00    78
## 2 2022-10-28 08:00:00    62
## 3 2022-10-28 16:00:00    41
## 4 2022-10-29 00:00:00    41
## 5 2022-10-29 08:00:00  1652
## 6 2022-10-29 16:00:00 21654
## 7 2022-10-30 00:00:00 19961
## 8 2022-10-30 08:00:00 10292
## 9 2022-10-30 16:00:00  4305
```

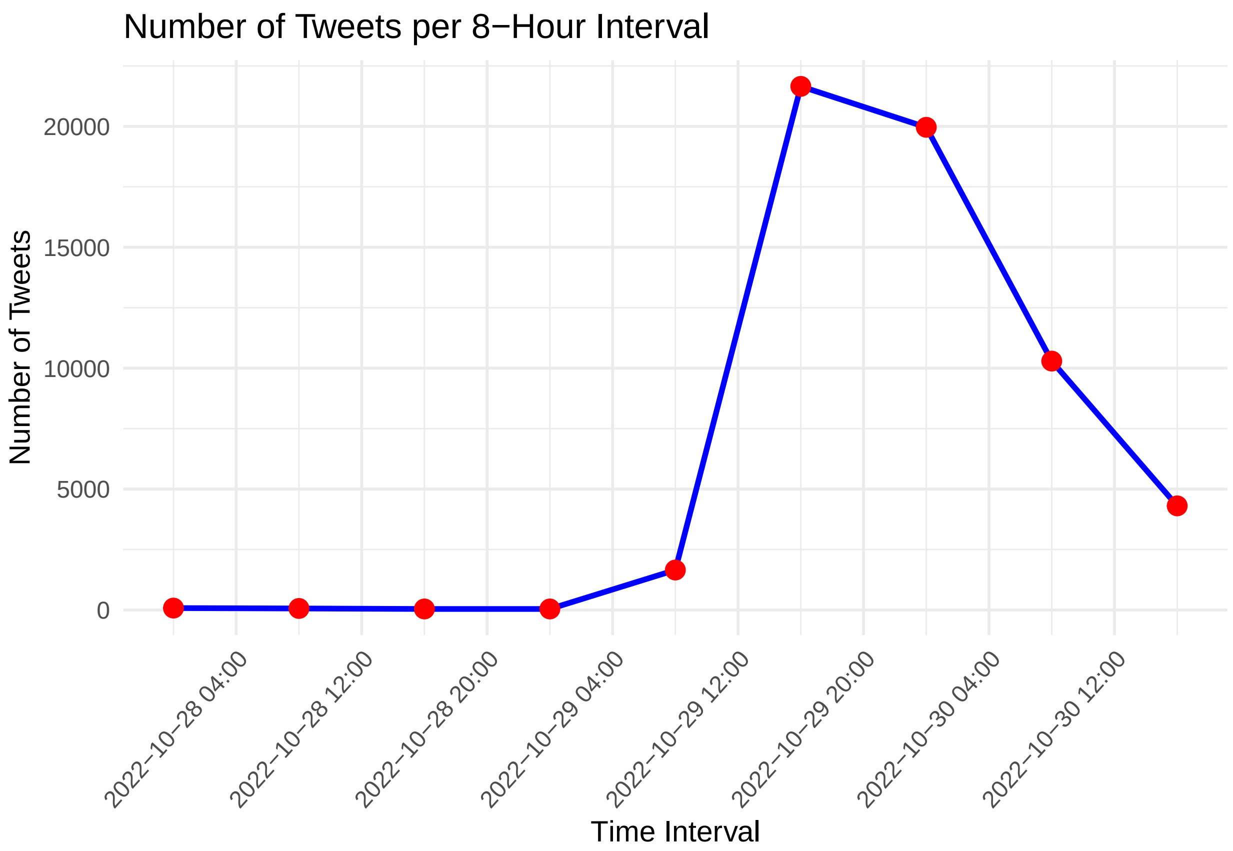
Saving the data as CSV

```
write.csv(hour_tweet, file = "Tweets_per_8_hours.csv", row.names = FALSE)
```

Plot of the data of based on each time interval

```
# Create a line plot
ggplot(hour_tweet, aes(x = interval, y = count)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
  labs(title = "Number of Tweets per 8-Hour Interval",
       x = "Time Interval",
       y = "Number of Tweets") +
  theme_minimal() +
  scale_x_datetime(date_labels = "%Y-%m-%d %H:%M", date_breaks = "8 hours") +
  theme(axis.text.x = element_text(angle = 48, hjust = 1))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
#The graph shows minimal tweet activity from October 28, 2022, until the early hours of October 29, 202
```

Cleaning of data in comments and make a wordcloud of it.

Accessing the data of the comments.

```
data_comments <- table(the_data$text)
head(data_comments)
```

```
##
## ''anti bullying'' on the bio and acting like the nugu oppa on your pfp won't get married too. LOL he
##
##                                                               'Death Alley' Itaewon \n.\nhttps://t.c
##
##        'Halloween crowd' reported 81 cases including difficulty breathing in #Itaewon"\n\n
##
##        'I was there when she said she couldn't breathe': Australian who saw friend crushed blames Se
##
##        'I was there when she said she couldn't breathe': Australian who saw friend crushed blames Se
##
##                                        'Itaewon crowd crush: 'I could not catch their pulse or breat
##
```

###Removing URL's that are on the comments

```
the_data$text <- gsub("http[s]?://\\S+", "", the_data$text)
head(the_data$text)
```

```
## [1] "A soldier angry at the support fund consolation money for the bereaved family of the Itaewon cr
## [2] "Nah this Itaewon tragedy really has me shaken up bc wth"
## [3] "@_J1N____ Pray for itaewon,  29.10.2022"
## [4] "TRANSLATION :\nSeoul residents lay flowers at a makeshift memorial near the site of the crush i
## [5] "The Itaewon stampede incident really caught me off guard. Makes me notice how important it is t
## [6] ""What to do about my child? What to do about my child?" Park Ga-young’s mother, Choi Seon-mi, s
```

###Setting the comments to lowercase, removing special characters and extra spaces.

```
clean_data <- the_data %>% mutate(text = str_to_lower(text), text = str_replace_all(text,"[^a-z ]", "")
head(clean_data$text)
```

```
## [1] "a soldier angry at the support fund consolation money for the bereaved family of the itaewon cr
## [2] "nah this itaewon tragedy really has me shaken up bc wth"
## [3] "jln pray for itaewon"
## [4] "translation seoul residents lay flowers at a makeshift memorial near the site of the crush in i
## [5] "the itaewon stampede incident really caught me off guard makes me notice how important it is to
## [6] "what to do about my child what to do about my child park gayoungs mother choi seonmi said as sh
```

Tokenize the Text

```
comm_text <- clean_data %>% unnest_tokens(word, text) %>% filter(!word %in% stop_words$word)
head(comm_text$word)
```

```
## [1] "soldier"     "angry"       "support"     "fund"        "consolation"
## [6] "money"
```

```
#This part makes the comment (text section of the csv) of the user to be word for word in order
#to foresee later the most searched word a user search.
```

Removing some unnecessary words manually.

```
custom_stop_words <- c("amp", "im", "whats", "omg", "yg", "ros",
                        "rn", "ive", "tw", "pls", "youre", "didnt", "shit",
                        "words", "bc", "wtf", "fucking", "yall", "fuck",
                        "africalink", "sm", "abt", "sa", "york", "damn",
                        "wont", "due", "wasnt", "dont", " " )
```

Filtering this unnecessary words to the data

```
comm_text <- comm_text %>%
  filter(!word %in% custom_stop_words)
```

Creating a term frequency table

```
word_frequnency <- comm_text %>% count(word, sort = TRUE)
head(word_frequnency)
```

```
##          word     n
## 1     itaewon 49271
## 2      people  7891
## 3    happened  6479
## 4   halloween  6430
## 5        pray  6428
## 6 condolences  6047
```

```
tail(word_frequnency)
```

```
##          word n
## 25697 zxarkin 1
## 25698  zxnhle 1
## 25699   zyada 1
## 25700  zylles 1
## 25701    zyrd 1
## 25702 zzrenra 1
```

Loading a sentiment using bing.

```
bing <- get_sentiments("bing")
comment_sentiment <- comm_text %>% inner_join(bing, by = "word")
head(comment_sentiment)
```

```
##   X    screenName             created
## 1 1      whourj31 2022-10-30 23:59:43
## 2 1      whourj31 2022-10-30 23:59:43
## 3 1      whourj31 2022-10-30 23:59:43
## 4 2       nnainot 2022-10-30 23:59:32
## 5 4 telehuntwatch 2022-10-30 23:59:28
## 6 4 telehuntwatch 2022-10-30 23:59:28
##                                                                        statusSource
## 1                <a href="https://www.fs-poster.com/" rel="nofollow">FS_Poster_App</a>
## 2                <a href="https://www.fs-poster.com/" rel="nofollow">FS_Poster_App</a>
## 3                <a href="https://www.fs-poster.com/" rel="nofollow">FS_Poster_App</a>
## 4 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 5                            <a href="https://ruprop.live" rel="nofollow">telehunt</a>
## 6                            <a href="https://ruprop.live" rel="nofollow">telehunt</a>
##      Created_At_Round tweetSource    word sentiment
## 1 2022-10-31 00:00:00      others   angry  negative
## 2 2022-10-31 00:00:00      others support  positive
## 3 2022-10-31 00:00:00      others   crush  negative
## 4 2022-10-31 00:00:00     android tragedy  negative
## 5 2022-10-31 00:00:00      others   crush  negative
## 6 2022-10-31 00:00:00      others    kill  negative
```

```
tail(comment_sentiment)
```

```
##           X    screenName             created
## 61152 58075     Ameraah_7 2022-10-28 00:56:26
## 61153 58075     Ameraah_7 2022-10-28 00:56:26
## 61154 58077    ivabangtan 2022-10-28 00:47:36
## 61155 58082 seokjlnswhore 2022-10-28 00:16:17
## 61156 58083 bjorktheworld 2022-10-28 00:15:24
## 61157 58086     N_ick_y28 2022-10-28 00:01:06
##                                                                            statusSource
## 61152   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 61153   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 61154   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 61155 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 61156   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 61157 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
##          Created_At_Round tweetSource         word sentiment
## 61152 2022-10-28 01:00:00      iphone         hate  negative
## 61153 2022-10-28 01:00:00      iphone unbelievable  negative
## 61154 2022-10-28 01:00:00      iphone        sweet  positive
## 61155 2022-10-28 00:00:00     android        sweet  positive
## 61156 2022-10-28 00:00:00      iphone       haters  negative
## 61157 2022-10-28 00:00:00     android        break  negative
```

Saving the data using RData

```
save(comment_sentiment, file = "Word_sentiments.RData")
```

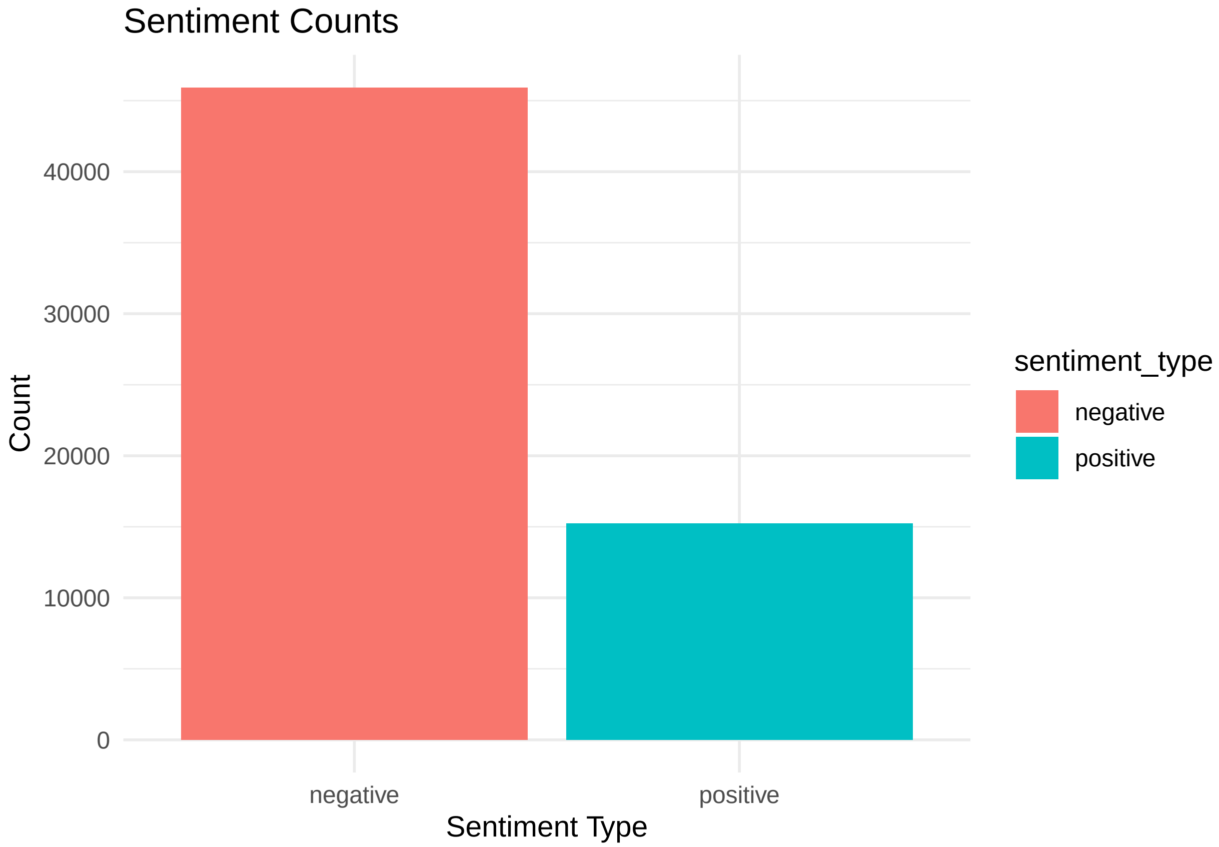
Analyzing the Sentiment by getting the total number of each sentiment

```
summary_sentiment <- comment_sentiment %>% count(sentiment) %>% pivot_wider(names_from = sentiment, valu
sentiment_summary <- summary_sentiment %>% mutate(net_sentiment = negative - positive)
sentiment_summary
```

```
## # A tibble: 1 x 3
##   negative positive net_sentiment
##      <int>    <int>         <int>
## 1    45927    15230         30697
```

```
sentiment_long <- sentiment_summary %>%
  pivot_longer(cols = c(positive, negative), names_to = "sentiment_type", values_to = "count")

ggplot(sentiment_long, aes(x = sentiment_type, y = count, fill = sentiment_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Sentiment Counts", x = "Sentiment Type", y = "Count") +
  theme_minimal()
```

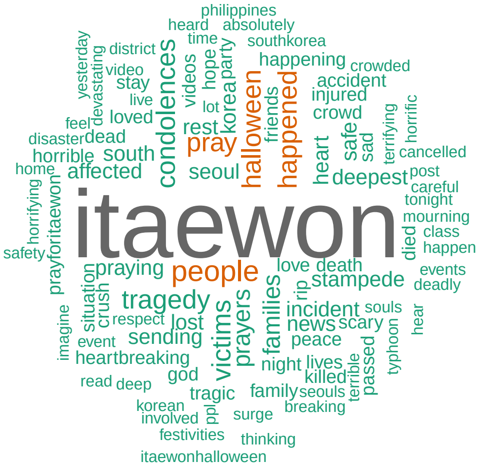


Description or Insights

```
#Based on the graph above, we can see that there are a lot of negative comments based on the tweets
# compare to the positive ones. Some of this tweets maybe because of their emotions in the incident
# in Itaewon. Some of them tweets maybe because they are angry at what happened feel sadness, surprised
# or fear.
```

Visualization of Result using WordCloud

```
set.seed(5000)
wordcloud(words = word_frequnency$word,
          freq = word_frequnency$n,
          min.freq = 1,
          max.words = 100,
          random.order = FALSE,
          rot.per = 0.35,
          scale = c(3, 0.5),
          colors = brewer.pal(8, "Dark2"))
```



Description or Insights

```
#The word cloud shows that 'itaewon' is the most frequent term in the tweets, standing out more than ot
```