SentimentAnalysis

Indhresh Achi

2025-04-13

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(ggplot2)
library(readr)
library(tidyr)
library(dplyr)
library(tidytext)
library(stringr)
library(textdata)
```

```
# Importing Dataset
```

```
sentiment <- read.csv("/Users/indhreshachi/Desktop/sentimentdataset.csv")
str(sentiment)
```

```
## 'data.frame':    732 obs. of  15 variables:
##  $ X         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Unnamed..0: int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Text      : chr  " Enjoying a beautiful day at the park!              "
" Traffic was terrible this morning.                 " " Just finished an
amazing workout! 💪                " " Excited about the upcoming weekend
getaway!         " ...
##  $ Sentiment : chr  " Positive  " " Negative  " " Positive  " " Positive
" ...
##  $ Timestamp : chr  "2023-01-15 12:30:00" "2023-01-15 08:45:00" "2023-01-
15 15:45:00" "2023-01-15 18:20:00" ...
##  $ User      : chr  " User123      " " CommuterX    " " FitnessFan   " "
AdventureX   " ...
##  $ Platform  : chr  " Twitter  " " Twitter  " " Instagram " " Facebook "
```

```
...
##  $ Hashtags  : chr  " #Nature #Park                                " "
#Traffic #Morning                        " " #Fitness #Workout
" " #Travel #Adventure                      " ...
##  $ Retweets  : num  15 5 20 8 12 25 10 15 30 18 ...
##  $ Likes     : num  30 10 40 15 25 50 20 30 60 35 ...
##  $ Country   : chr  " USA      " " Canada   " " USA        " " UK       "
...
##  $ Year      : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
##  $ Month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day       : int  15 15 15 15 15 16 16 16 17 17 ...
##  $ Hour      : int  12 8 15 18 19 9 14 19 8 12 ...
```

```
colnames(sentiment)
```

```
##  [1] "X"          "Unnamed..0" "Text"       "Sentiment"  "Timestamp"
##  [6] "User"       "Platform"   "Hashtags"   "Retweets"   "Likes"
## [11] "Country"    "Year"       "Month"      "Day"        "Hour"
```

```
# Data Wrangling
```

```
head(sentiment)
```

```
##   X Unnamed..0                                                Text
Sentiment
## 1 0          0  Enjoying a beautiful day at the park!
Positive
## 2 1          1  Traffic was terrible this morning.
Negative
## 3 2          2  Just finished an amazing workout! 💪
Positive
## 4 3          3  Excited about the upcoming weekend getaway!
Positive
## 5 4          4  Trying out a new recipe for dinner tonight.
Neutral
## 6 5          5  Feeling grateful for the little things in life.
Positive
##             Timestamp           User    Platform
## 1 2023-01-15 12:30:00  User123         Twitter
## 2 2023-01-15 08:45:00  CommuterX       Twitter
## 3 2023-01-15 15:45:00  FitnessFan     Instagram
## 4 2023-01-15 18:20:00  AdventureX      Facebook
## 5 2023-01-15 19:55:00  ChefCook       Instagram
## 6 2023-01-16 09:10:00  GratitudeNow    Twitter
##                                    Hashtags Retweets Likes      Country
Year
## 1  #Nature #Park                                  15    30    USA
2023
## 2  #Traffic #Morning                               5    10    Canada
2023
## 3  #Fitness #Workout                              20    40  USA
```

```
2023
## 4  #Travel #Adventure                                  8    15    UK
2023
## 5  #Cooking #Food                                     12    25   Australia
2023
## 6    #Gratitude #PositiveVibes                        25    50    India
2023
##   Month Day Hour
## 1     1  15   12
## 2     1  15    8
## 3     1  15   15
## 4     1  15   18
## 5     1  15   19
## 6     1  16    9
```

```
sentiment$`Unnamed..0` <- NULL
sentiment$X <-  NULL
sentiment <-  subset(sentiment, select=-c(Year, Month, Day, Hour))

sentiment$Platform = trimws(sentiment$Platform)
sentiment$Sentiment = trimws(sentiment$Sentiment)
sentiment$Country = trimws(sentiment$Country)
sentiment$Hashtags = trimws(sentiment$Hashtags)

glimpse(sentiment)
```

```
## Rows: 732
## Columns: 9
## $ Text      <chr> " Enjoying a beautiful day at the park!              ",
" Tr…
## $ Sentiment <chr> "Positive", "Negative", "Positive", "Positive",
"Neutral", "…
## $ Timestamp <chr> "2023-01-15 12:30:00", "2023-01-15 08:45:00", "2023-01-
15 15…
## $ User      <chr> " User123      ", " CommuterX    ", " FitnessFan   ", "
Adve…
## $ Platform  <chr> "Twitter", "Twitter", "Instagram", "Facebook",
"Instagram", …
## $ Hashtags  <chr> "#Nature #Park", "#Traffic #Morning", "#Fitness
#Workout", "…
## $ Retweets  <dbl> 15, 5, 20, 8, 12, 25, 10, 15, 30, 18, 22, 7, 12, 28, 15,
20,…
## $ Likes     <dbl> 30, 10, 40, 15, 25, 50, 20, 30, 60, 35, 45, 15, 25, 55,
30, …
## $ Country   <chr> "USA", "Canada", "USA", "UK", "Australia", "India",
"Canada"…
```

```
unique(sentiment$Platform)
```

```
## [1] "Twitter"   "Instagram" "Facebook"
```

```r
summary(sentiment$Likes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   34.75   43.00   42.90   50.00   80.00
```

```r
sd(sentiment$Likes)
```

```
## [1] 14.08985
```

```r
summary(sentiment$Retweets)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.75   22.00   21.51   25.00   40.00
```

```r
sd(sentiment$Retweets)
```

```
## [1] 7.061286
```

```r
head(unique(sentiment$Sentiment))
```

```
## [1] "Positive" "Negative" "Neutral"  "Anger"    "Fear"     "Sadness"
```

```r
glimpse(sentiment)
```

```
## Rows: 732
## Columns: 9
## $ Text      <chr> " Enjoying a beautiful day at the park!              ",
" Tr…
## $ Sentiment <chr> "Positive", "Negative", "Positive", "Positive",
"Neutral", "…
## $ Timestamp <chr> "2023-01-15 12:30:00", "2023-01-15 08:45:00", "2023-01-
15 15…
## $ User      <chr> " User123      ", " CommuterX    ", " FitnessFan   ", "
Adve…
## $ Platform  <chr> "Twitter", "Twitter", "Instagram", "Facebook",
"Instagram", …
## $ Hashtags  <chr> "#Nature #Park", "#Traffic #Morning", "#Fitness
#Workout", "…
## $ Retweets  <dbl> 15, 5, 20, 8, 12, 25, 10, 15, 30, 18, 22, 7, 12, 28, 15,
20,…
## $ Likes     <dbl> 30, 10, 40, 15, 25, 50, 20, 30, 60, 35, 45, 15, 25, 55,
30, …
## $ Country   <chr> "USA", "Canada", "USA", "UK", "Australia", "India",
"Canada"…
```

```r
sentiment %>%
  group_by(Platform) %>%
  summarize(avg_retweet = mean(Retweets)) %>%
  select(Platform, avg_retweet)
```

```
## # A tibble: 3 × 2
##   Platform  avg_retweet
##   <chr>           <dbl>
```

```
## 1 Facebook         21.0
## 2 Instagram        22.6
## 3 Twitter          20.9
```

```
sentiment %>%
  group_by(Platform) %>%
  summarize(avg_retweet = mean(Retweets)) %>%
  select(Platform, avg_retweet)
```

```
## # A tibble: 3 × 2
##   Platform  avg_retweet
##   <chr>           <dbl>
## 1 Facebook         21.0
## 2 Instagram        22.6
## 3 Twitter          20.9
```

```
unique(sentiment$Platform)
```

```
## [1] "Twitter"   "Instagram" "Facebook"
```

```
sentiment %>%
  group_by(Platform) %>%
  summarize(avg_likes = mean(Likes)) %>%
  select(Platform, avg_likes)
```

```
## # A tibble: 3 × 2
##   Platform  avg_likes
##   <chr>         <dbl>
## 1 Facebook       41.9
## 2 Instagram      45.1
## 3 Twitter        41.6
```

```
facebook <- sentiment[sentiment$Platform == "Facebook", ]
```

```
instagram <- sentiment[sentiment$Platform == "Instagram", ]
```

```
usa <- sentiment[sentiment$Country == "USA", ]
head(usa$Text,10)
```

```
##  [1] " Enjoying a beautiful day at the park!              "
##  [2] " Just finished an amazing workout! 💪               "
##  [3] " The new movie release is a must-watch!             "
##  [4] " Political discussions heating up on the timeline.  "
##  [5] " Just published a new blog post. Check it out!      "
##  [6] " New year, new fitness goals! 💪                    "
##  [7] " Reflecting on the past and looking ahead.          "
##  [8] " Attending a virtual conference on AI.              "
##  [9] " Winter blues got me feeling low.                   "
## [10] " Exploring the world of virtual reality.            "
```

```
unique(sentiment$Country)
```

```
##  [1] "USA"            "Canada"         "UK"             "Australia"
##  [5] "India"          "France"         "Brazil"         "Japan"
##  [9] "Greece"         "Germany"        "Sweden"         "Italy"
## [13] "Netherlands"    "South Africa"   "Spain"          "Portugal"
## [17] "Switzerland"    "Austria"        "Belgium"        "Denmark"
## [21] "Czech Republic" "Jordan"         "Peru"           "Maldives"
## [25] "China"          "Cambodia"       "Norway"         "Colombia"
## [29] "Ireland"        "Jamaica"        "Kenya"          "Scotland"
## [33] "Thailand"
```

```
india <- sentiment[sentiment$Country == "India",]
head(india$Text,10)
```

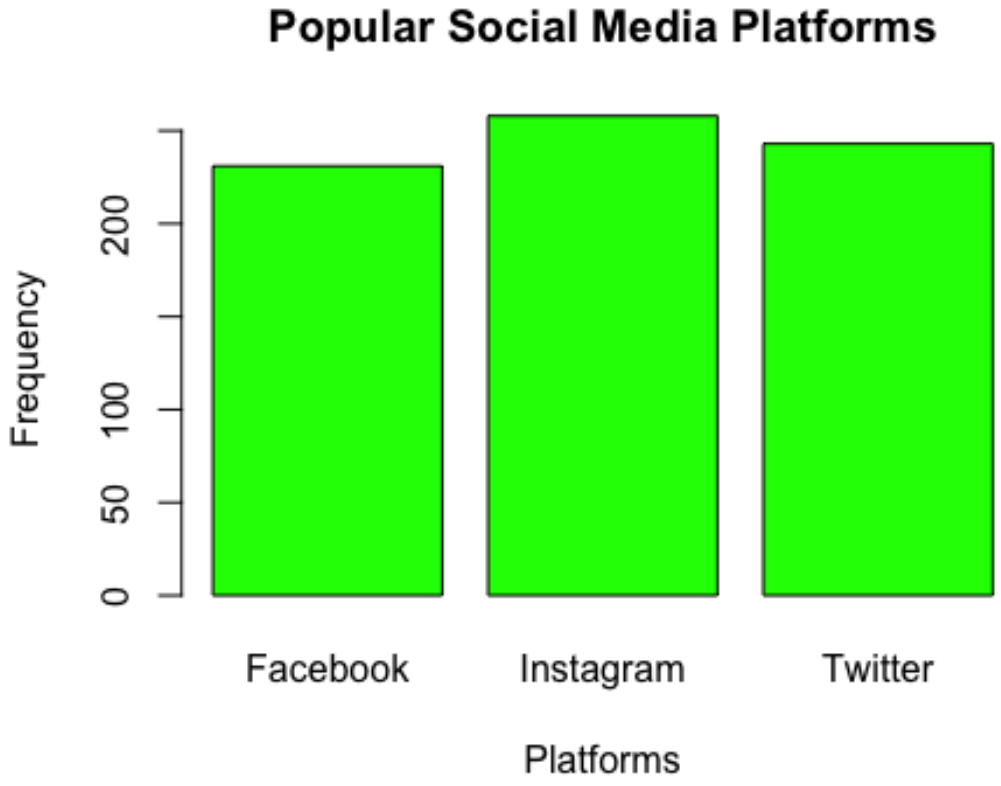
```
##  [1] " Feeling grateful for the little things in life.    "
##  [2] " Technology is changing the way we live.            "
##  [3] " Sipping coffee and enjoying a good book.           "
##  [4] " Learning a new language for personal growth.       "
##  [5] " Enjoying a cup of tea and watching the sunset.     "
##  [6] " Practicing mindfulness with meditation.            "
##  [7] " Feeling accomplished after a productive day.       "
##  [8] " Attending a virtual reality meetup.                "
##  [9] " Heartbroken after hearing the news about a natural disaster. "
## [10] " Disappointed with the service at a local restaurant.    "
```

```
japan <- sentiment[sentiment$Country == "Japan",]
head(japan$Text, 10)
```

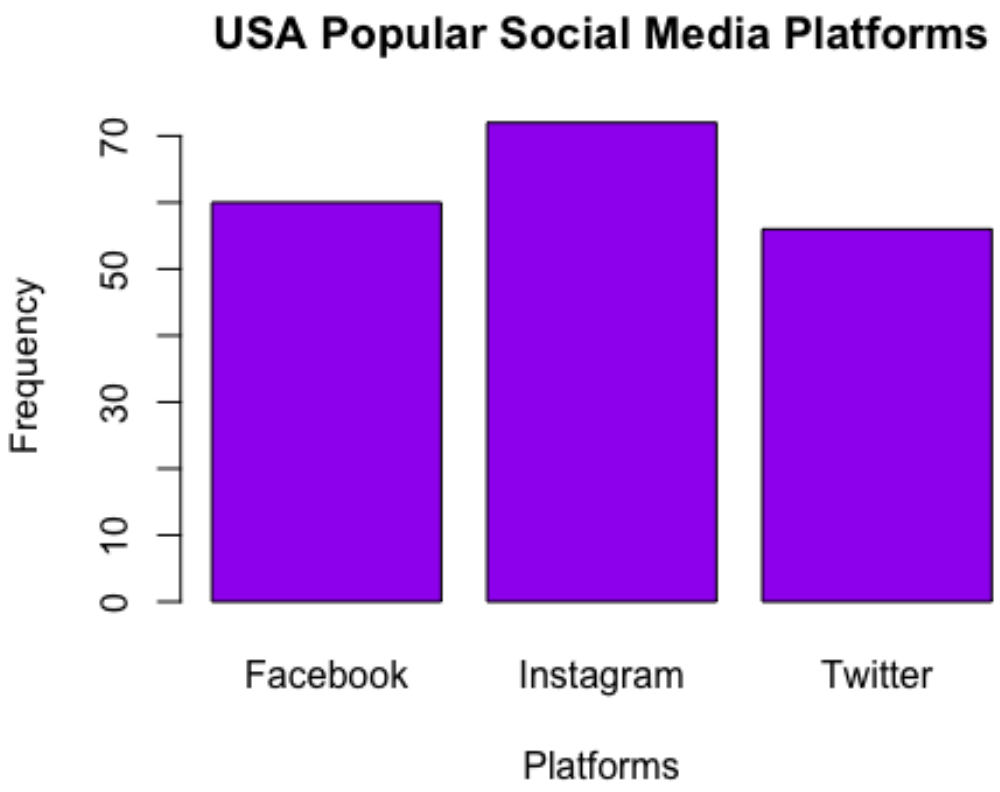
```
##  [1] "Bittersweet emotions arise while bidding farewell to a dear friend.
"
##  [2] "Giggles and joy echo in the air during a children's playdate.      "
##  [3] "Joyful laughter resonates through a lively summer carnival.
"
##  [4] "Imbued with gratitude for the simple pleasure of a warm cup of tea.
"
##  [5] "Isolation deepens, an emotional winter where warmth is but a distant
memory. "
##  [6] "Avoiding the shards of shattered dreams, walking the tightrope of
resilience. "
##  [7] "In the void of heartache, echoes of a love song play, each note a
pang of longing. "
##  [8] "Overflowing with joy, a cup of laughter shared with friends, a
moment cherished. "
##  [9] "In the celebration of success, fireworks of accomplishment light up
the night sky of triumph. "
## [10] "Wandering through the historical streets of Kyoto, each step a
journey into the heart of Japan's traditions. "
```

```
# Data Visualization

platform_freq = table(sentiment$Platform)
usa_platforms = table(usa$Platform)
```

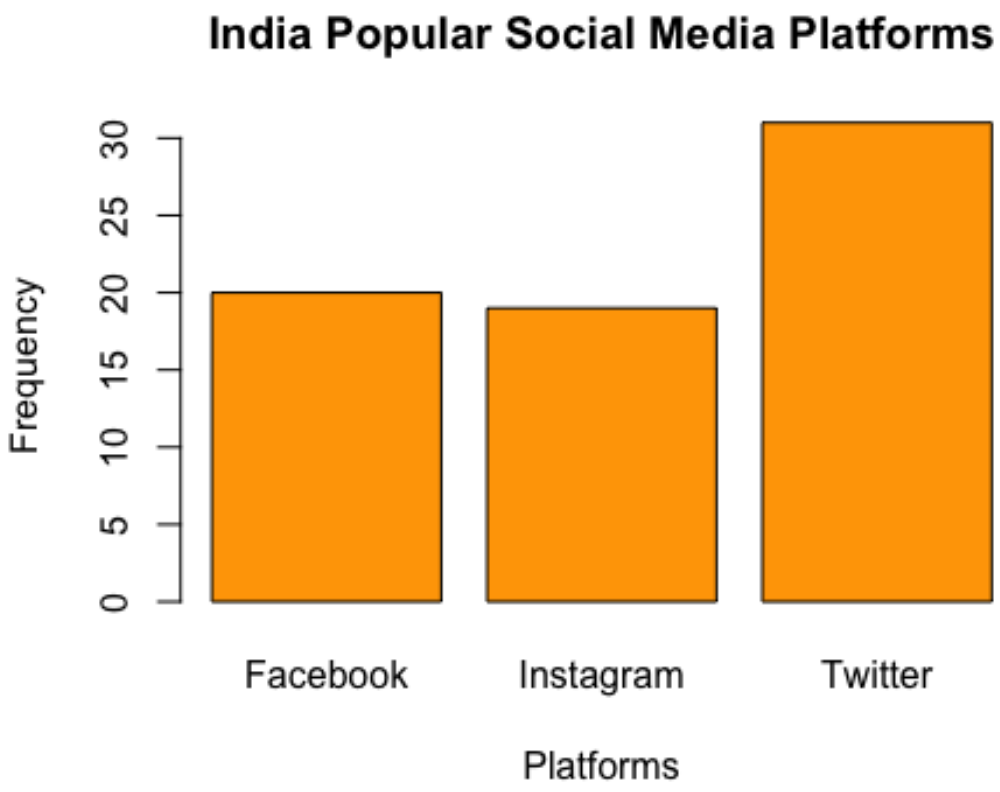
```r
india_platforms = table(india$Platform)
japan_platforms = table(japan$Platform)
barplot(platform_freq,
        main="Popular Social Media Platforms",
        xlab="Platforms",
        ylab="Frequency", col="green")
```



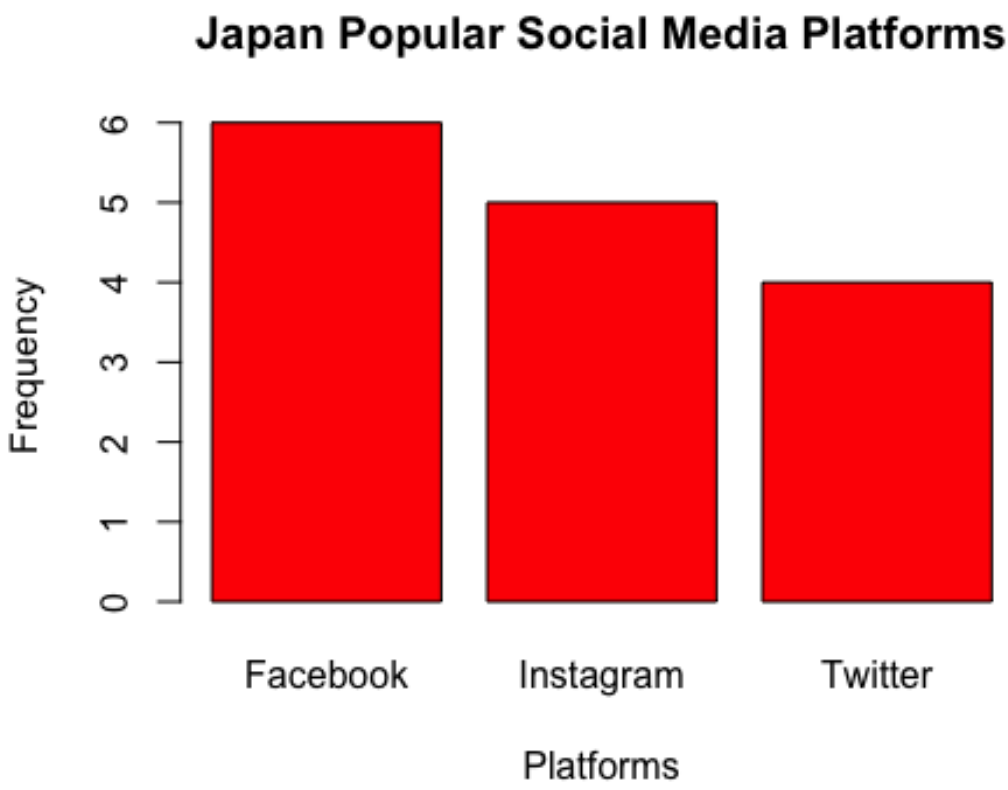

```r
barplot(usa_platforms,
        main="USA Popular Social Media Platforms",
        xlab="Platforms",
        ylab="Frequency", col="purple")
```




```r
barplot(india_platforms,
        main="India Popular Social Media Platforms",
        xlab="Platforms",
        ylab="Frequency", col="orange")
```




```r
barplot(japan_platforms,
        main="Japan Popular Social Media Platforms",
        xlab="Platforms",
        ylab="Frequency", col="red")
```




```
head(sentiment$Platform, 3)
```

```
## [1] "Twitter"   "Twitter"   "Instagram"
```

```
afinn = get_sentiments(lexicon="afinn")

sentimentValues_df <- sentiment %>%
  mutate(index=row_number()) %>%
  unnest_tokens(word, Sentiment) %>%
  inner_join(afinn) %>%
  group_by(index) %>%
  summarise(sentimentVal = sum(value)) %>%
  left_join(
    sentiment%>%
      mutate(index=row_number())
  )
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(index)`
```

```
textValues_df <- sentimentValues_df %>%
  mutate(index=row_number()) %>%
  unnest_tokens(word, Text) %>%
  inner_join(afinn) %>%
```

```
  group_by(index) %>%
  summarise(textVal = sum(value))%>%
  left_join(
    sentimentValues_df%>%
      mutate(index=row_number())
  )

## Joining with `by = join_by(word)`
## Joining with `by = join_by(index)`

encoded <- textValues_df %>%
  mutate(index=row_number()) %>%
  unnest_tokens(word, Hashtags) %>%
  inner_join(afinn) %>%
  group_by(index) %>%
  summarise(HashtagsVal = sum(value)) %>%
  left_join(
    textValues_df%>%
      mutate(index=row_number())
  )

## Joining with `by = join_by(word)`
## Joining with `by = join_by(index)`

encoded$sentimentVal

##   [1]  2  2  2  2  2 -3 -2 -3  3  3  3  3 -2  1 -2  3 -2 -3 -2 -3  3  3  3
3  3
##  [26] -2  1  3 -2  3 -2  3 -2  3  3  2  3  3  2  3  3  3  2  3  3  2  3 -3
-2 -2
##  [51] -2 -2 -1 -2 -3 -3 -2 -2 -2 -2 -1 -2 -3 -3 -2 -2 -2 -2 -2 -2 -2 -2 -2
-2 -2
##  [76]  2  2  3  2  2  2  2  2  2  2  2  2  2  2  2  3  2  2  2  2  2  2  2
2  2
## [101]  2  2  3  2  2  2  2  2  2  2  2  2 -3  1 -2 -2 -2 -2 -2 -2 -2 -2 -2
-2 -2
## [126] -3  1 -2 -2 -2 -2 -2 -2 -2  3  4  2  1  3  3  1  2  2  2  3  3 -1  2
2  2
## [151]  3  2  2  2  3  3  1  3 -2 -3 -3 -2 -2 -2 -2 -3 -2 -2 -1 -2 -2 -2 -2
-3 -2
## [176] -3 -2 -2 -3 -3 -3 -2 -3 -3 -2 -3 -3  3  3  3  3  3  3  2  2  3  3  2
3  1
## [201]  3  3  2  3  3  2  3  2  4 -2 -2 -2  2 -1 -1 -3  2  3 -2

encoded$textVal

##   [1]   4   3   3   3   3  -2  -1  -3   9   6   5   3  -2   2  -2   3  -4
-5
##  [19]  -2  -3   3   4   6   6   3  -2  -1   3  -4   6  -1   3  -1   4   3
2
##  [37]   3   3   2   3   3   3   3   4   6   4   3  -5   2  -5  -5  -2  -1
```

```
-3
##  [55]  -8  -1  -4  -2  -3  -4  -1  -2  -3  -3  -2  -2  -5  -1  -3  -2  -2
-2
##  [73]  -2  -2  -1   5   3   2   2   2   5   4   7   2  -1   4   4   6   7
0
##  [91]   3   3   5   3   2   4  -1   4   4   6   7   0   3   3   5   3   2
4
## [109]  -1   4   4   6 -10   5  -4  -4   3  -4  -4   0  -3  -4  -6  -1  -2
-3
## [127]   3 -10  -3  -2  -5  -2   3  -3   2   6   4   3   5   3   3   5   0
3
## [145]   5   3  -4   4   2   2   3   4   4   2   6   1   5   3   0  -3  -2
-4
## [163]  -5  -7  -1  -3  -2  -7  -4  -3  -4  -2  -4  -2  -6  -4   0  -5  -5
-3
## [181]   5  -2  -2  -4  -4  -1  -2   5   4   6   7   5   3   4   6   3   3
2
## [199]   3  -2   3   4   2   4   5   5  11   8   4  -4  -3  -4   0  -3  -5
-1
## [217]   2   4  -1
```

```
encoded$HashtagsVal
```

```
##   [1]  1  2  2  4  2 -5 -2 -3  3  3  3  3 -2  3  2  3 -4 -3 -2 -3  3  3  3
3  3
##  [26] -2  1  3 -4  3 -2  3  0  3  3  4  3  3  2  3  3  3  2  3  3  2  3 -5
-5 -4
##  [51] -5 -4  0 -2 -5 -5 -4 -2 -2 -2  1 -2 -3 -3 -2 -2 -2 -2 -2 -2 -2 -2 -2
-2 -2
##  [76]  2  2  3  2  2  2  2  2  2  2  2  2  2  2  2  3  2  2  2  2  2  2  2
2  2
## [101]  2  2  3  2  2  2  2  2  2  2  2  2 -3  1 -2 -2 -2 -2 -2 -2 -2 -2 -2
-2 -2
## [126] -3  1 -2 -2 -2 -2 -2 -2 -2  3  4  2  1  3  3  1  2  2  2  3  3 -1  2
2  2
## [151]  3  2  2  2  3  3  1  3 -2 -3 -3 -2 -2 -2 -2 -3 -2 -2 -1 -2 -2 -2 -2
-3 -2
## [176] -3 -2 -2 -3 -3 -3 -2 -3 -3 -2 -3 -3  3  3  3  3  3  3  2  2  3  3  2
3  1
## [201]  3  3  2  3  3  2  3  2  4 -2 -2 -2  2 -1 -1 -3  2  3 -2
```

```
names(encoded)
```

```
##  [1] "index"        "HashtagsVal"  "textVal"      "sentimentVal" "Text"
##  [6] "Sentiment"    "Timestamp"    "User"         "Platform"
"Hashtags"
## [11] "Retweets"     "Likes"        "Country"
```

```
cluster_df <- subset(encoded, select=
                       -c(index,
                          Timestamp, User,Hashtags,
```

```
                          Text, Sentiment))
cluster_df %>%
  group_by(Platform) %>%
  summarize(sentimentSum = sum(sentimentVal))
```

```
## # A tibble: 3 × 2
##   Platform  sentimentSum
##   <chr>            <dbl>
## 1 Facebook            18
## 2 Instagram           51
## 3 Twitter              5
```

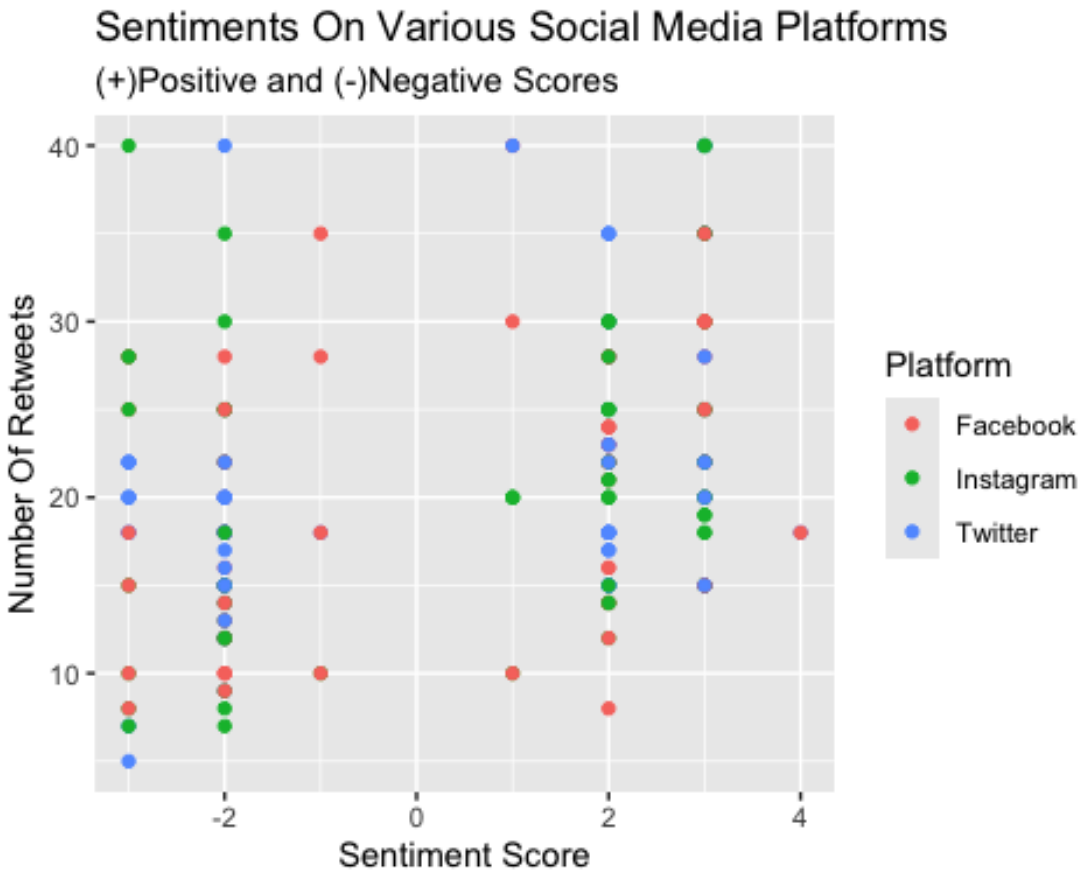
```
cluster_df %>%
  group_by(Platform) %>%
  summarize(textSum = sum(textVal))
```

```
## # A tibble: 3 × 2
##   Platform  textSum
##   <chr>       <dbl>
## 1 Facebook       35
## 2 Instagram      76
## 3 Twitter        43
```

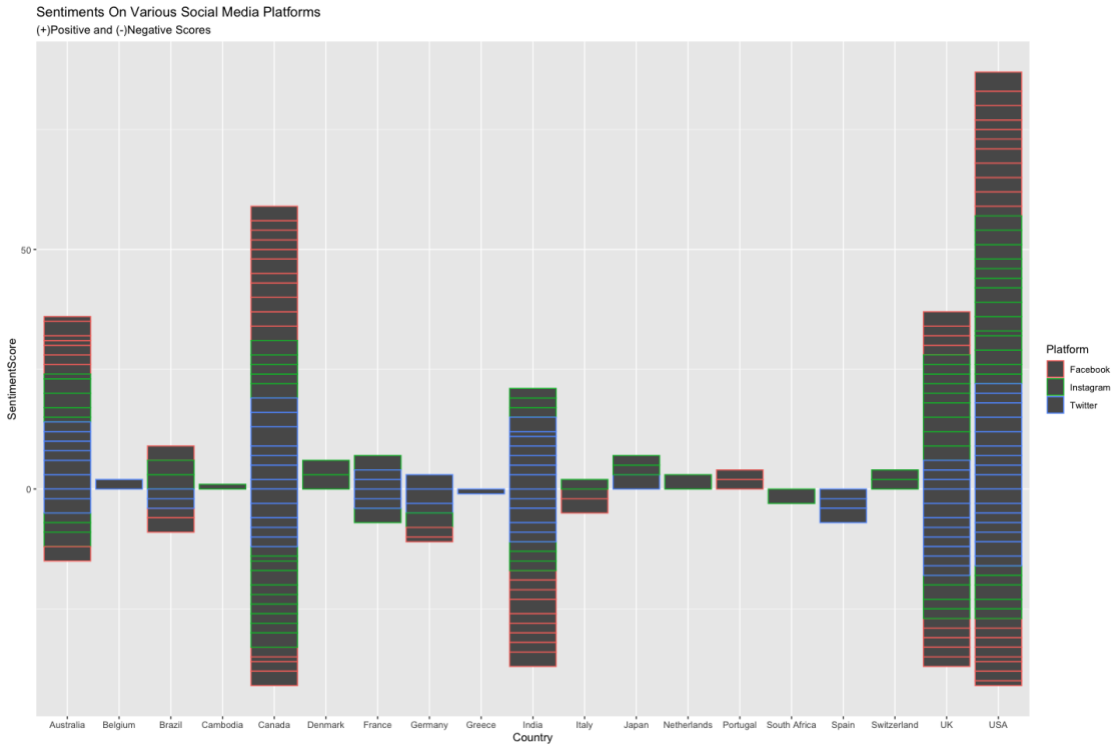
```
cluster_df %>%
  group_by(Platform) %>%
  summarize(hashtagSum = sum(HashtagsVal))
```

```
## # A tibble: 3 × 2
##   Platform  hashtagSum
##   <chr>          <dbl>
## 1 Facebook          16
## 2 Instagram         47
## 3 Twitter            1
```

```
ggplot(data=cluster_df, mapping=aes(x=sentimentVal, y=Retweets))+
  geom_point(aes(color=Platform))+
  ggtitle(label="Sentiments On Various Social Media Platforms",
          subtitle="(+)Positive and (-)Negative Scores ")+
  xlab("Sentiment Score")+
  ylab("Number Of Retweets")
```




```
ggplot(data=cluster_df, mapping=aes(x=Country, y=sentimentVal))+
  geom_col(aes(color=Platform))+
  ggtitle(label="Sentiments On Various Social Media Platforms",
          subtitle="(+)Positive and (-)Negative Scores ")+
  xlab("Country")+
  ylab("SentimentScore")
```

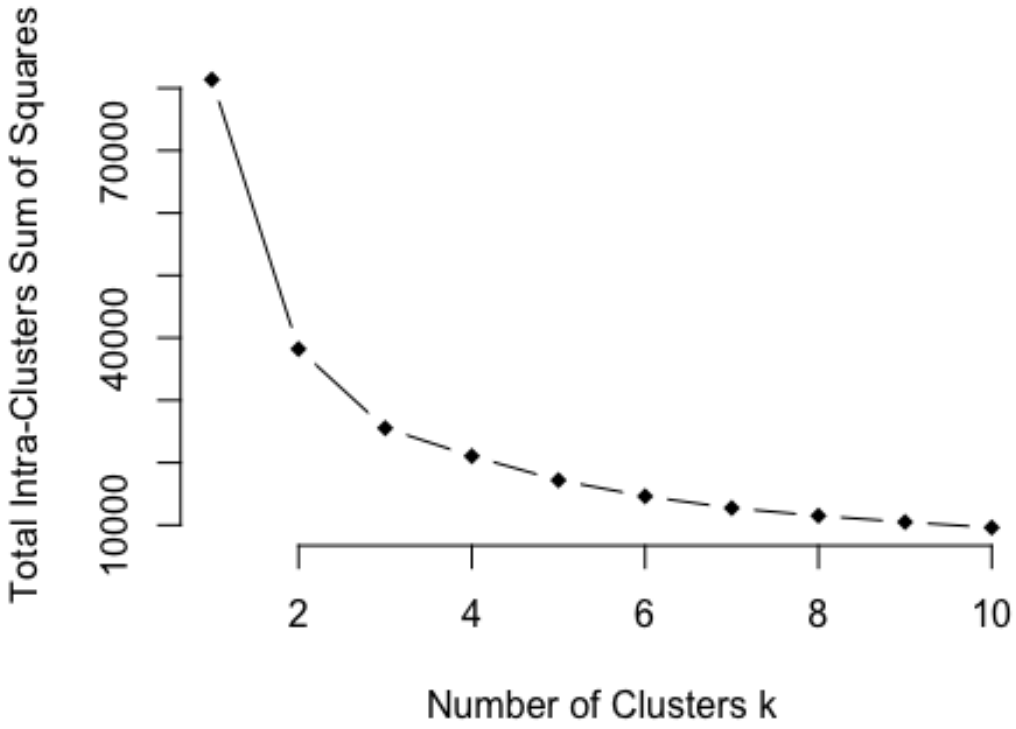



```
# Machine Learning

## K-Means Clustering

sentiment$Country <- unclass(factor(sentiment$Country))
sentiment$Platform <- unclass(factor(sentiment$Platform))
library(purrr)
set.seed(1234)
glimpse(sentiment)

## Rows: 732
## Columns: 9
## $ Text      <chr> " Enjoying a beautiful day at the park!              ",
" Tr…
## $ Sentiment <chr> "Positive", "Negative", "Positive", "Positive",
"Neutral", "…
## $ Timestamp <chr> "2023-01-15 12:30:00", "2023-01-15 08:45:00", "2023-01-
15 15…
## $ User      <chr> " User123      ", " CommuterX    ", " FitnessFan   ", "
Adve…
## $ Platform  <int> 3, 3, 2, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1,
2, …
## $ Hashtags  <chr> "#Nature #Park", "#Traffic #Morning", "#Fitness
#Workout", "…
## $ Retweets  <dbl> 15, 5, 20, 8, 12, 25, 10, 15, 30, 18, 22, 7, 12, 28, 15,
20,…
## $ Likes     <dbl> 30, 10, 40, 15, 25, 50, 20, 30, 60, 35, 45, 15, 25, 55,
```

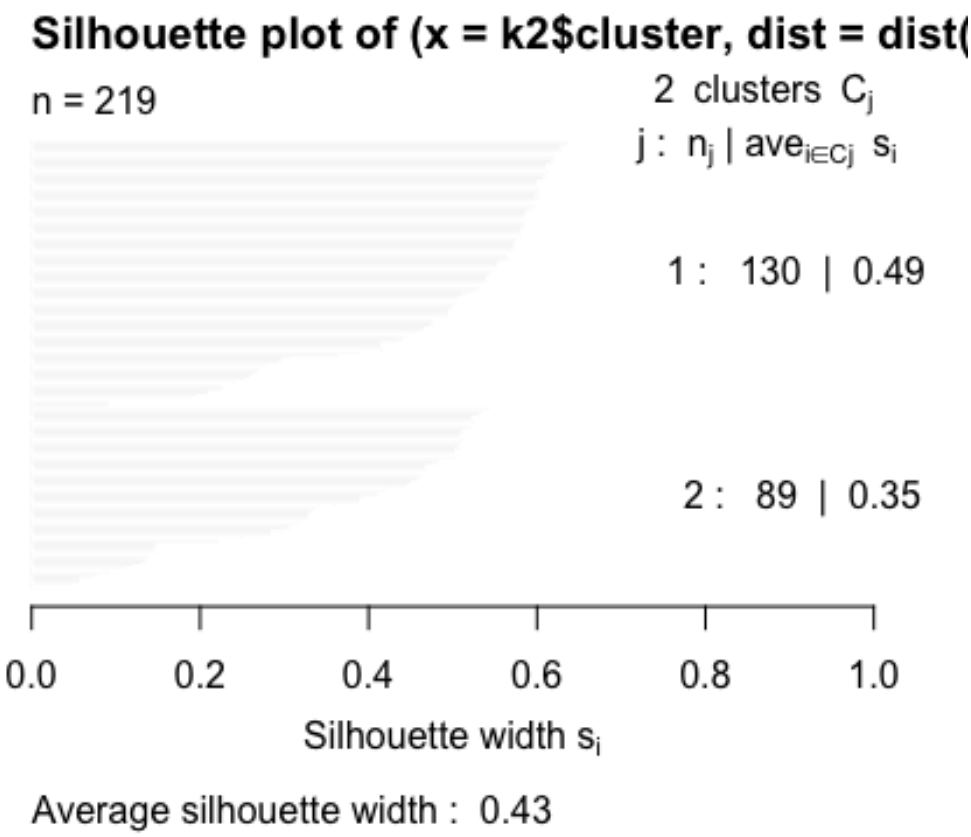
```
30, …
## $ Country   <int> 33, 6, 33, 32, 1, 14, 6, 33, 33, 1, 33, 6, 32, 33, 14,
33, 6…

names(cluster_df)

## [1] "HashtagsVal"  "textVal"      "sentimentVal" "Platform"     "Retweets"
## [6] "Likes"        "Country"

cluster_df$Platform <- as.numeric(as.factor(cluster_df$Platform))
cluster_df$Country <- as.numeric(as.factor(cluster_df$Country))
```

Elbow Method

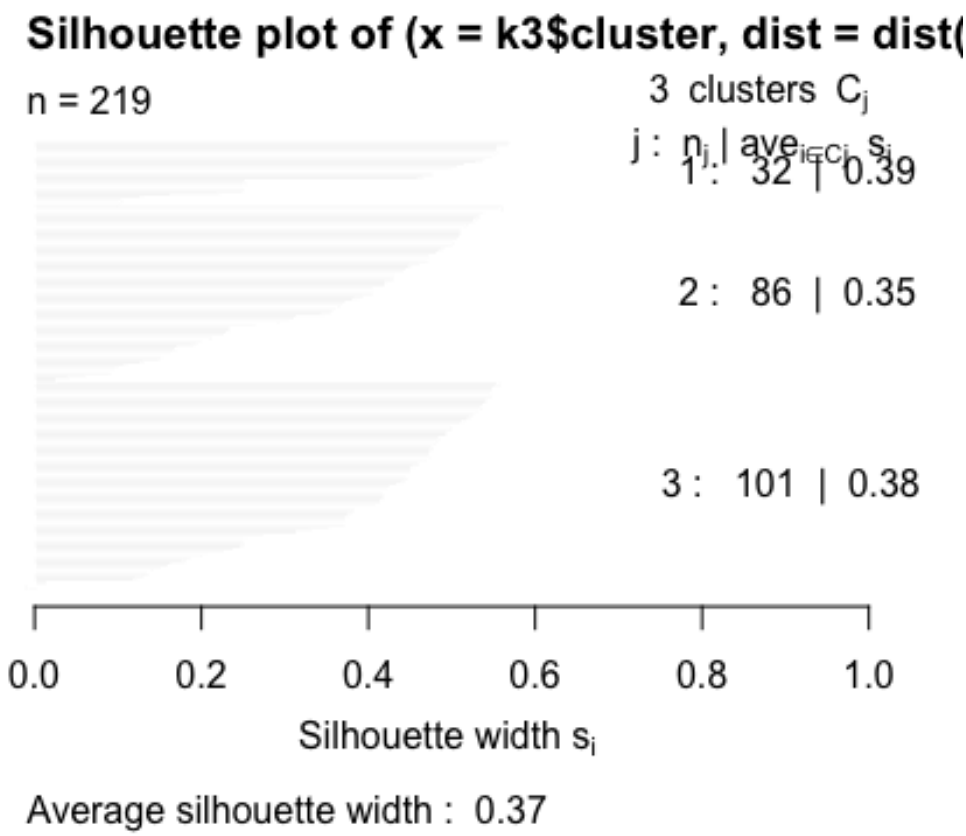
```
head(cluster_df$Platform, 3)

## [1] 2 1 2

iss <- function(k) {
  kmeans(cluster_df, k,
         iter.max = 100, nstart=100, algorithm = "Lloyd")$tot.withinss
}

k.values <- 1:10
iss_values <- map_dbl(k.values, iss)

plot(k.values, iss_values, type="b", pch=18, frame=FALSE,
     xlab="Number of Clusters k",
     ylab="Total Intra-Clusters Sum of Squares")
```



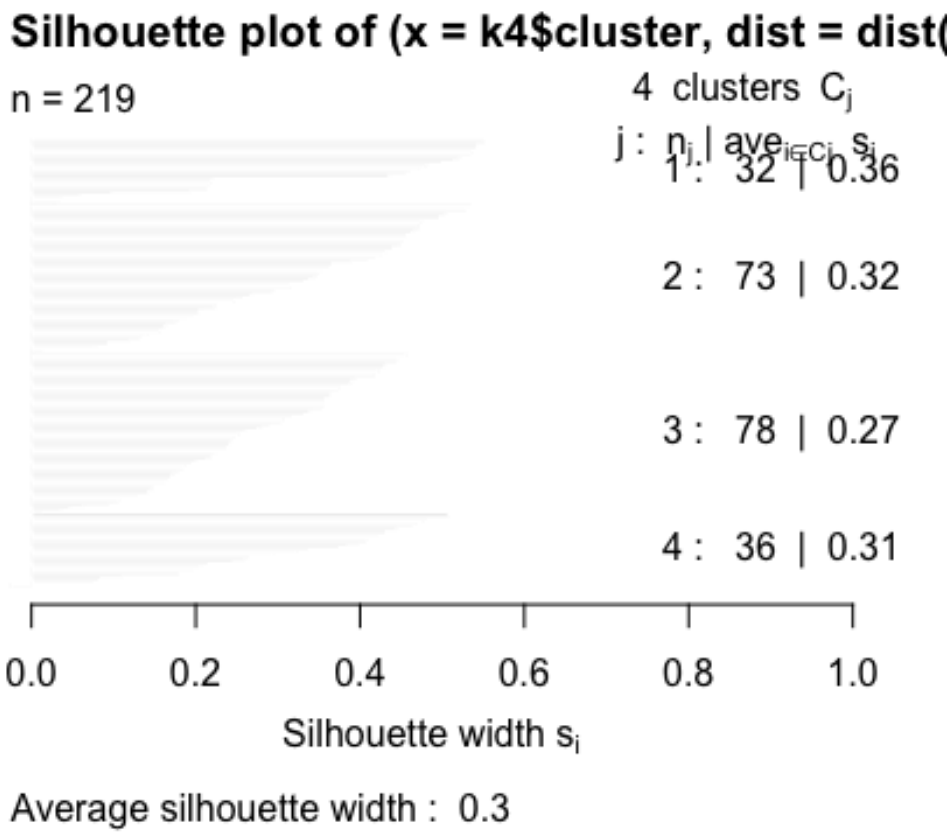

```
### Silhouette Method


library(cluster)
library(grid)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

k2 <- kmeans(cluster_df,
             2, iter.max = 100, nstart = 50, algorithm="Lloyd")
s2 <- plot(silhouette(
  k2$cluster, dist(cluster_df, "euclidean")))
```



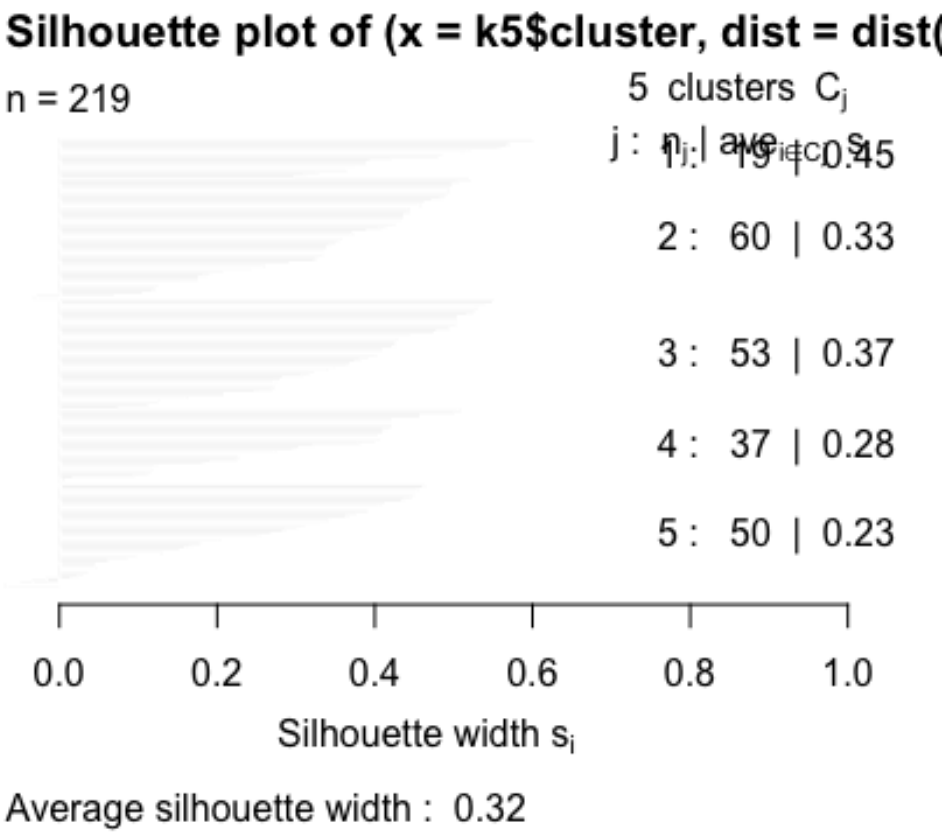

```
k3 <- kmeans(cluster_df,
             3, iter.max = 100, nstart = 50, algorithm="Lloyd")
s3 <- plot(silhouette(
  k3$cluster, dist(cluster_df, "euclidean")))
```



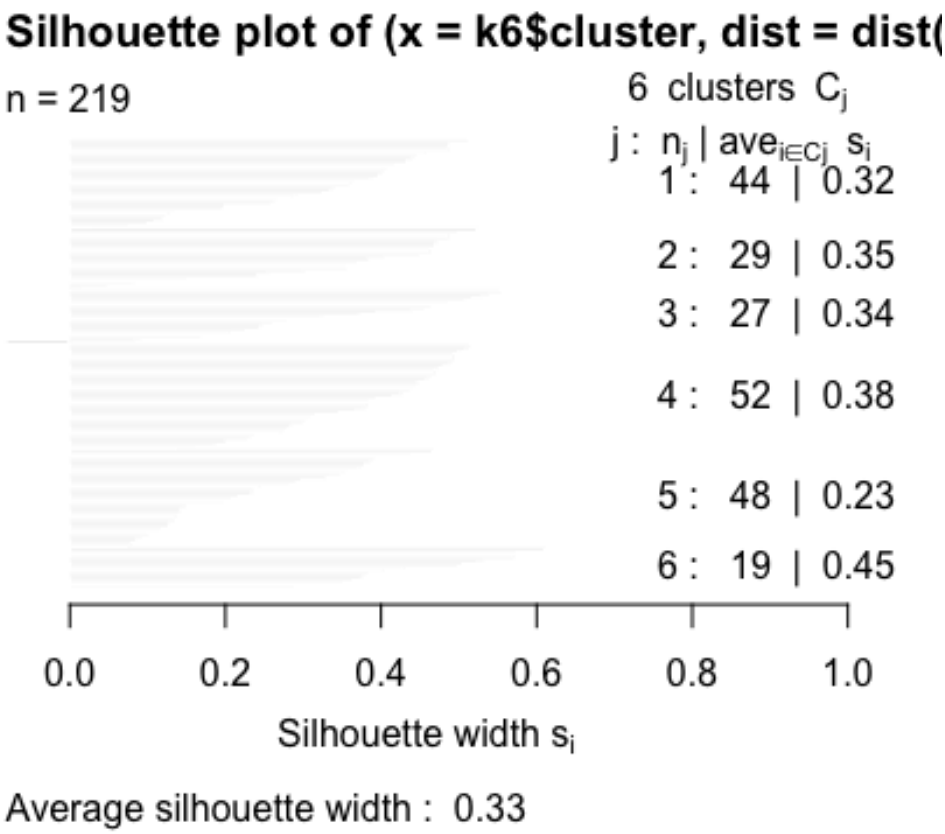

```
k4 <- kmeans(cluster_df,
             4, iter.max = 100, nstart = 50, algorithm="Lloyd")
s4 <- plot(silhouette(
  k4$cluster, dist(cluster_df, "euclidean")))
```



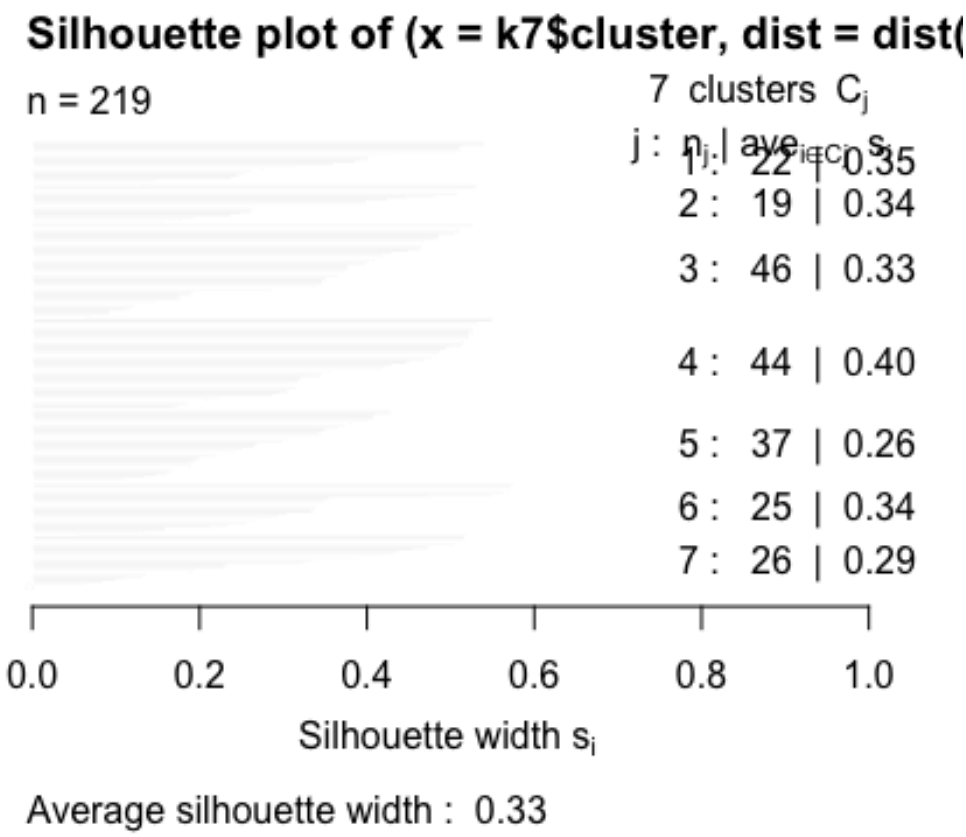

```r
k5 <- kmeans(cluster_df,
             5, iter.max = 100, nstart = 50, algorithm="Lloyd")
s5 <- plot(silhouette(
  k5$cluster, dist(cluster_df, "euclidean")))
```



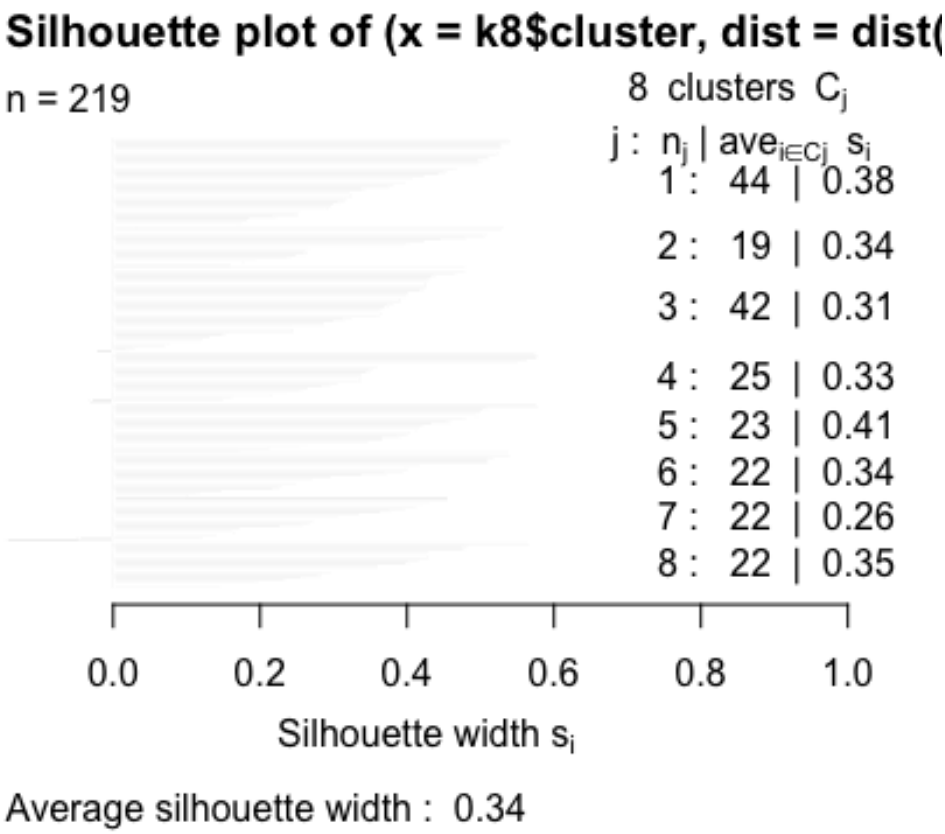

```r
k6 <- kmeans(cluster_df,
             6, iter.max = 100, nstart = 50, algorithm="Lloyd")
s6 <- plot(silhouette(
  k6$cluster, dist(cluster_df, "euclidean")))
```



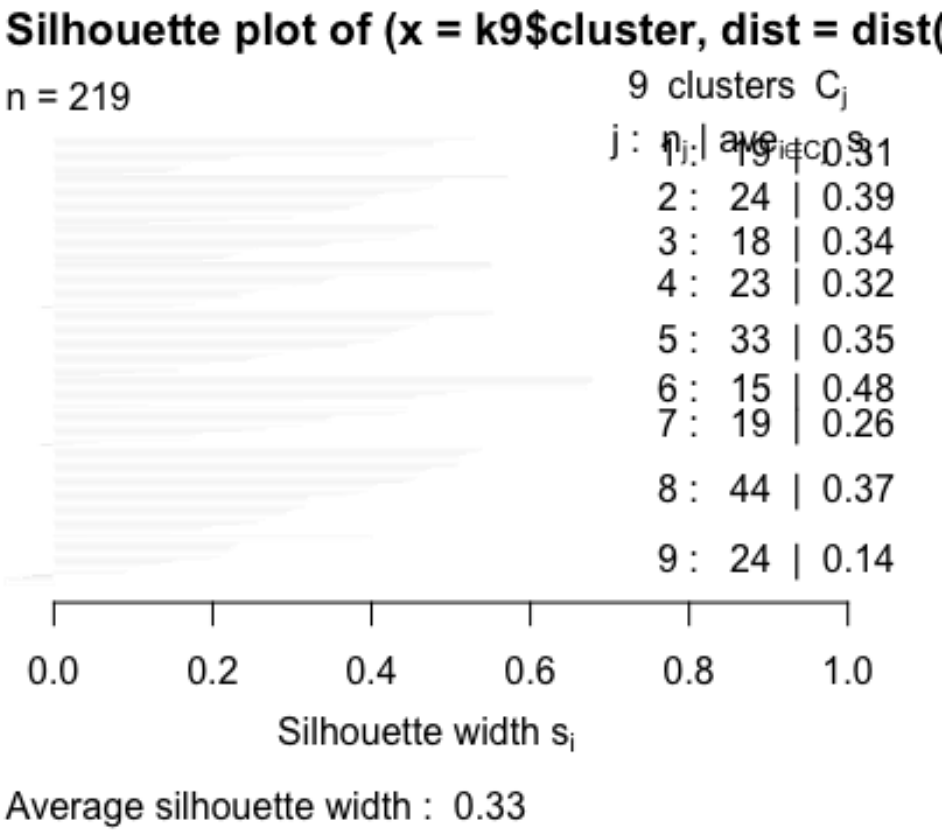

```
k7 <- kmeans(cluster_df,
             7, iter.max = 100, nstart = 50, algorithm="Lloyd")
s7 <- plot(silhouette(
  k7$cluster, dist(cluster_df, "euclidean")))
```



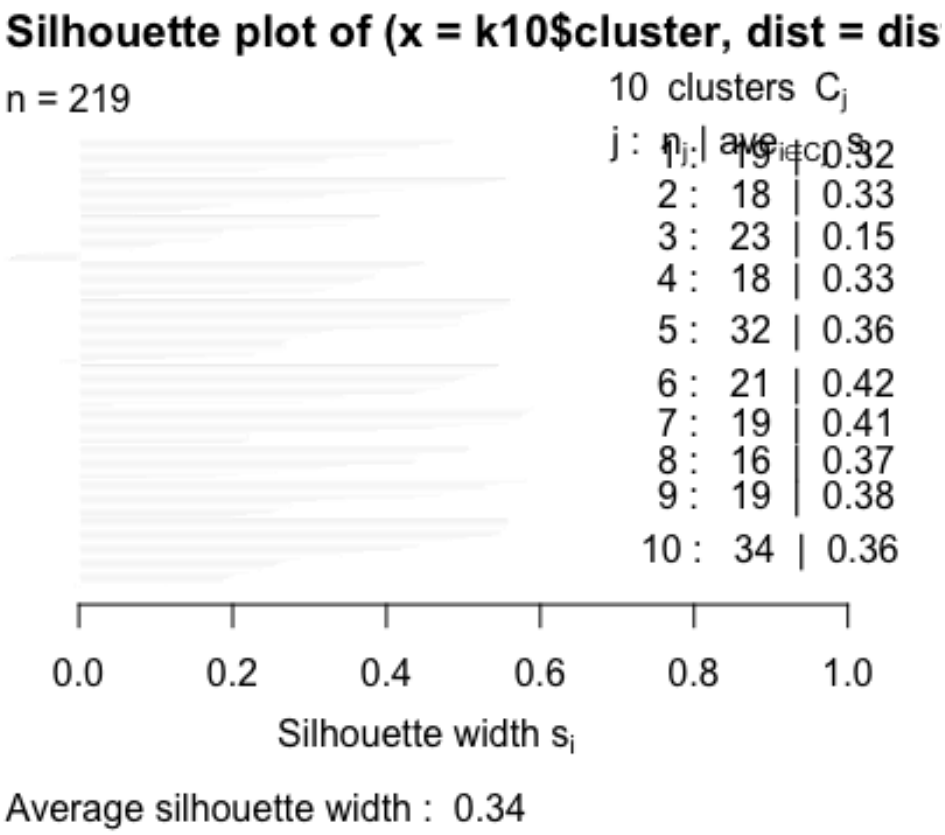

```r
k8 <- kmeans(cluster_df,
             8, iter.max = 100, nstart = 50, algorithm="Lloyd")
s8 <- plot(silhouette(
  k8$cluster, dist(cluster_df, "euclidean")))
```



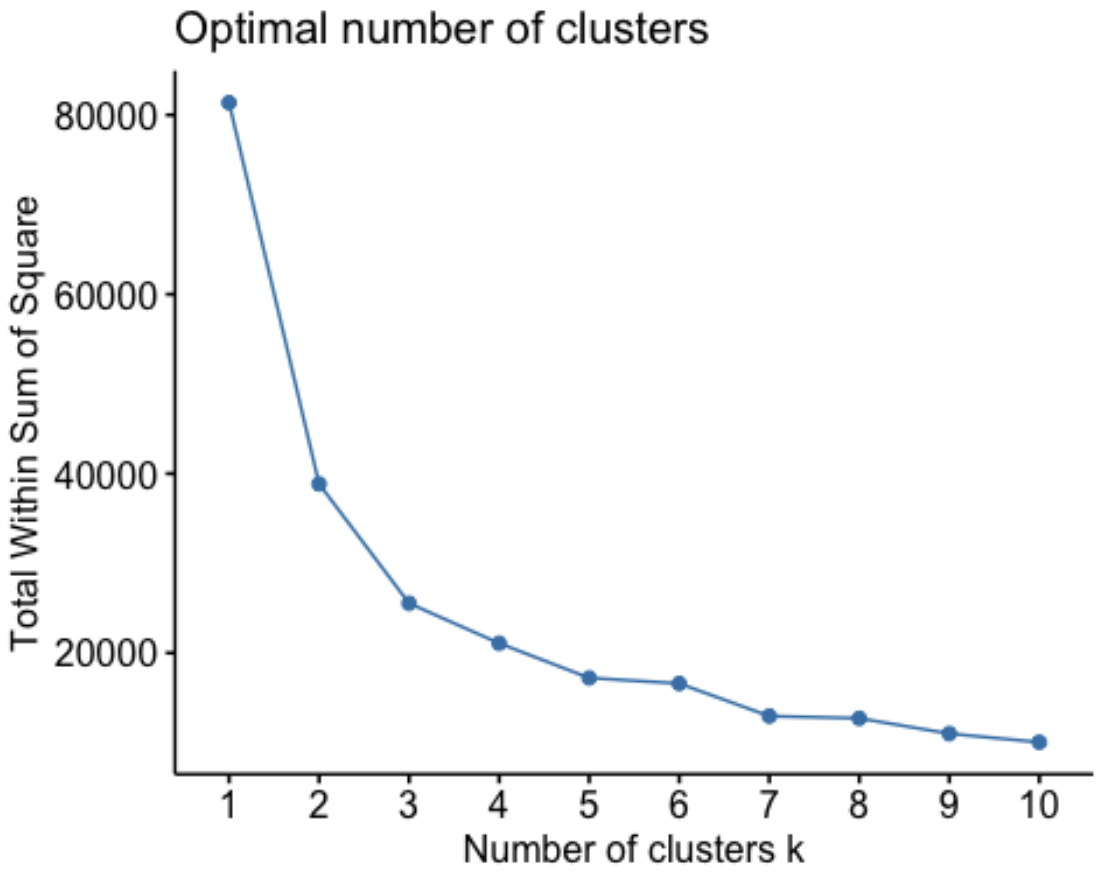

```r
k9 <- kmeans(cluster_df,
             9, iter.max = 100, nstart = 50, algorithm="Lloyd")
s9 <- plot(silhouette(
  k9$cluster, dist(cluster_df, "euclidean")))
```



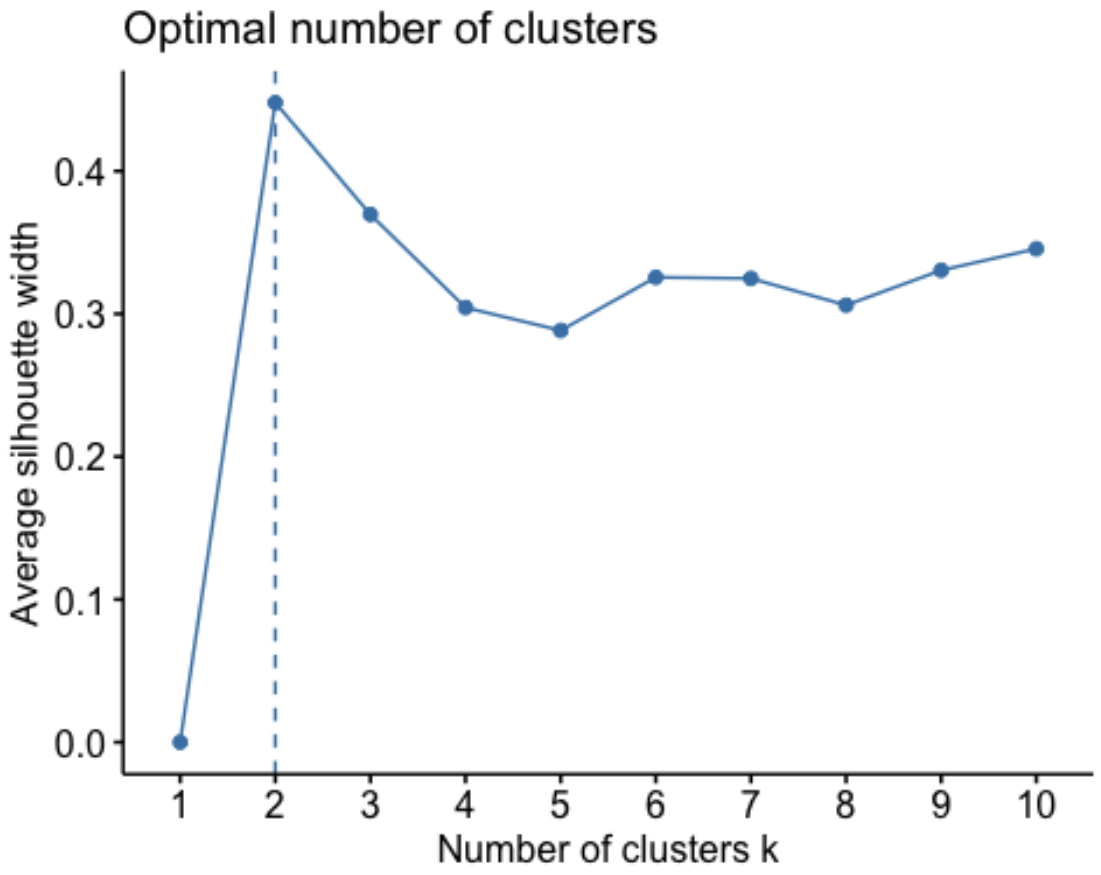

```
k10 <- kmeans(cluster_df,
              10, iter.max = 100, nstart = 50, algorithm="Lloyd")
s10 <- plot(silhouette(
  k10$cluster, dist(cluster_df, "euclidean")))
```



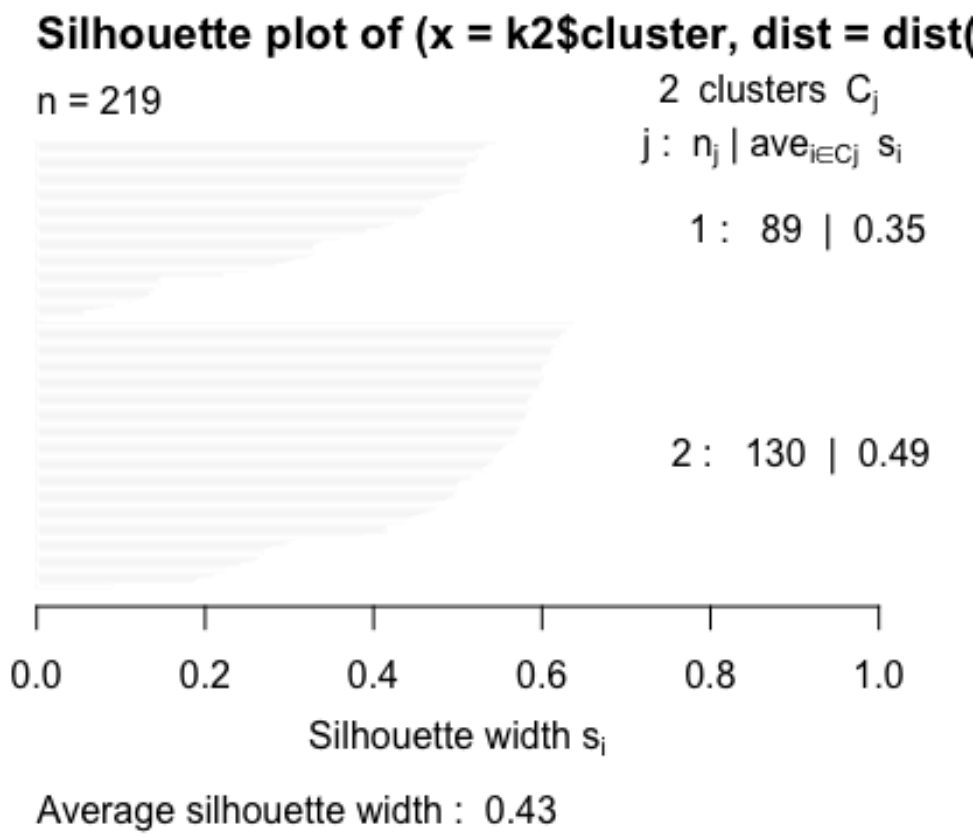

```
library(ggplot2)
library(NbClust)
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

fviz_nbclust(cluster_df, kmeans, method="wss")
```




```r
fviz_nbclust(cluster_df, kmeans, method="silhouette")
```

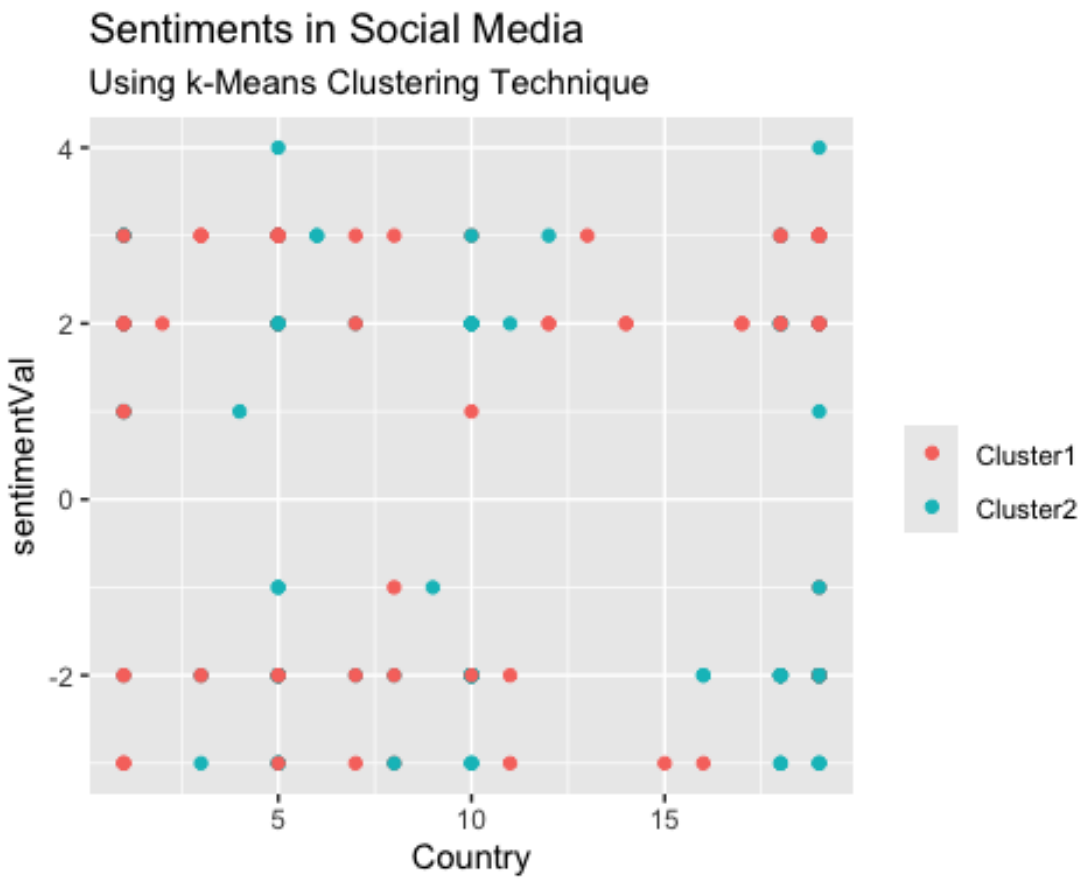



```
### Gap Statistic Method

k2 <- kmeans(cluster_df,
             2, iter.max = 100, nstart = 50, algorithm="Lloyd")
s2 <- plot(silhouette(
  k2$cluster, dist(cluster_df, "euclidean")))
```

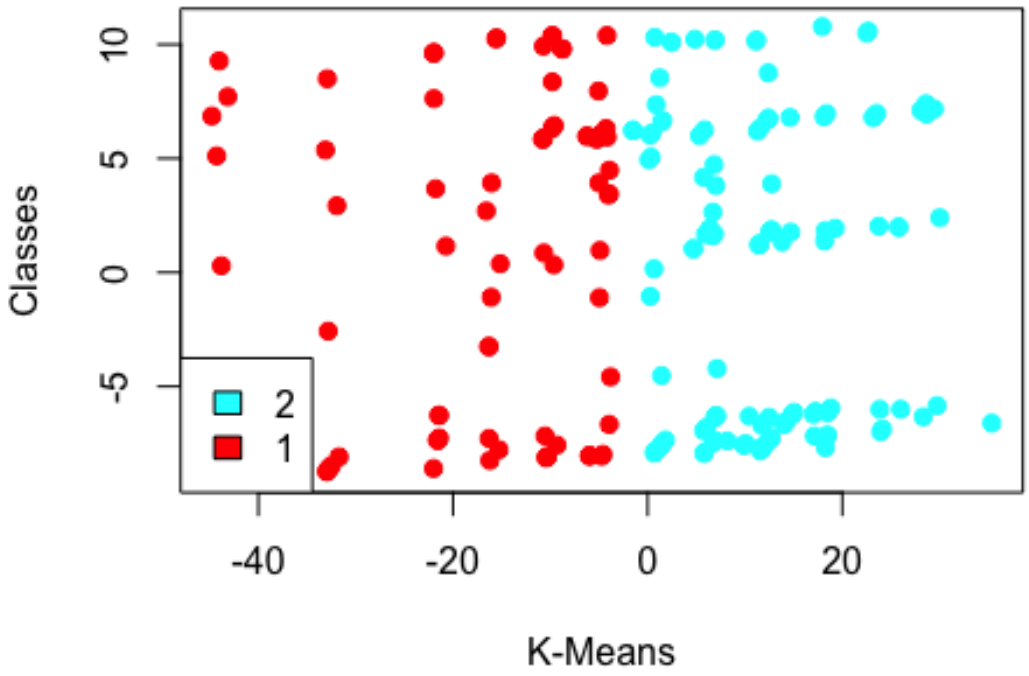



```
#### Principle Cluster Analysis

pcclust <- prcomp(cluster_df)

summary(pcclust)

## Importance of components:
##                             PC1    PC2     PC3     PC4     PC5    PC6
PC7
## Standard deviation     17.4175 6.7540 4.59809 1.50138 0.80573 0.4312
0.24111
## Proportion of Variance  0.8127 0.1222 0.05664 0.00604 0.00174 0.0005
0.00016
## Cumulative Proportion   0.8127 0.9349 0.99157 0.99761 0.99935 0.9998
1.00000

pcclust$rotation

##                        PC1           PC2           PC3            PC4
PC5
## HashtagsVal  -0.055857077 -0.025318804  0.465174934  0.5910181326
0.0390097547
## textVal      -0.073993342 -0.041345997  0.758629021 -0.6440115720 -
```

```
0.0345561787
## sentimentVal -0.050083358 -0.020391341  0.440808132  0.4820668655
0.0272986516
## Platform     -0.001152129  0.000752123 -0.004151975 -0.0586504823
0.9982075465
## Retweets     -0.445484867 -0.012617856 -0.048766647  0.0002681241
0.0086555927
## Likes        -0.888717939 -0.015443022 -0.094194037 -0.0106971590 -
0.0067282935
## Country       0.024878915 -0.998416445 -0.050145174  0.0019542337
0.0006308872
##                      PC6          PC7
## HashtagsVal  -0.654635784 -0.021886581
## textVal      -0.036616330  0.003422332
## sentimentVal  0.754480729  0.019727928
## Platform      0.003996634 -0.010330049
## Retweets     -0.025267486  0.893472104
## Likes         0.014414021 -0.447997721
## Country       0.002807336 -0.004359573

names(cluster_df)

## [1] "HashtagsVal"  "textVal"      "sentimentVal" "Platform"     "Retweets"
## [6] "Likes"        "Country"

### Platforms Encoding
# Twitter: 3
# Instagram: 2
# Facebook: 1
names(cluster_df)

## [1] "HashtagsVal"  "textVal"      "sentimentVal" "Platform"     "Retweets"
## [6] "Likes"        "Country"

ggplot(cluster_df, aes(x=Country, y=sentimentVal)) +
  geom_point(stat='identity', aes(color=as.factor(k2$cluster))) +
  scale_color_discrete(name=' ',
                       breaks=c("1","2"),
                       labels=c("Cluster1", "Cluster2"))+
  ggtitle("Sentiments in Social Media",
          subtitle="Using k-Means Clustering Technique")
```




```
### Country Encoding
# USA: 33
# China: 7
# India: 14
# Japan: 18

kcols = function(vec){
  cols=rainbow(length(unique(vec)))
  return (cols[as.numeric(as.factor(vec))])
}

digCluster <- k2$cluster;

dignm <- as.character(digCluster)

plot(pcclust$x, col=kcols(digCluster), pch=19, xlab="K-Means",
ylab="Classes")
legend("bottomleft", unique(dignm), fill=unique(kcols(digCluster)))
```


Classes
10
5
0
-5
2
1
-40
-20
0
20
K-Means