

Emotions in social media

Marc Cervera Rosell

2024-08-15

1. Load dataset

Dataset link

```
tryCatch({  
  data <- read.csv("train.csv", header = TRUE)  
  print("File read successfully")  
}, error = function(e) {  
  cat("ERROR when loading the dataset", conditionMessage(e), "\n")  
})
```

```
## [1] "File read successfully"
```

2. Preprocessing data

2.1 Delete blank lines (if needed):

```
cat("Rows before:", nrow(data), "\n")
```

```
## Rows before: 2004
```

```
data <- data[rowSums(is.na(data) | data == "") != ncol(data), ]  
cat("Rows after:", nrow(data))
```

```
## Rows after: 1000
```

2.2 Check variable types and column names

```
columns <- names(data)  
types <- sapply(data, class)  
for (i in seq_along(columns)) {  
  cat("Column name:", columns[i], " Type:", types[i], "\n")  
}
```

```
## Column name: User_ID Type: integer  
## Column name: Age Type: integer  
## Column name: Gender Type: character  
## Column name: Platform Type: character  
## Column name: Daily_Usage_Time..minutes. Type: integer  
## Column name: Posts_Per_Day Type: integer  
## Column name: Likes_Received_Per_Day Type: integer  
## Column name: Comments_Received_Per_Day Type: integer  
## Column name: Messages_Sent_Per_Day Type: integer
```

```
## Column name: Dominant_Emotion Type: character
```

Transformations:

- Column “Age” will become an integer
- Column “Daily_Usage_Time..minutes” will be renamed as “Minutes_Per_Day”

```
data_transformed <- transform(data,
                              Age = as.integer(Age))
colnames(data_transformed)[colnames(data_transformed) ==
                           "Daily_Usage_Time..minutes."] <- "Minutes_Per_Day"
```

```
types <- sapply(data_transformed, class)
for (i in seq_along(columns)) {
  cat("Column name:", columns[i], " Type:", types[i], "\n")
}
```

```
## Column name: User_ID Type: integer
## Column name: Age Type: integer
## Column name: Gender Type: character
## Column name: Platform Type: character
## Column name: Daily_Usage_Time..minutes. Type: integer
## Column name: Posts_Per_Day Type: integer
## Column name: Likes_Received_Per_Day Type: integer
## Column name: Comments_Received_Per_Day Type: integer
## Column name: Messages_Sent_Per_Day Type: integer
## Column name: Dominant_Emotion Type: character
```

2.3 Check if there's NA values

```
any(is.na(data_transformed))
```

```
## [1] FALSE
```

3. Descriptive analysis and inferential

3.1 Data distribution per gender, platform, age and dominant emotion

3.1.1 Data distribution per gender

```
genders <- unique(data_transformed$Gender)
print(genders)
```

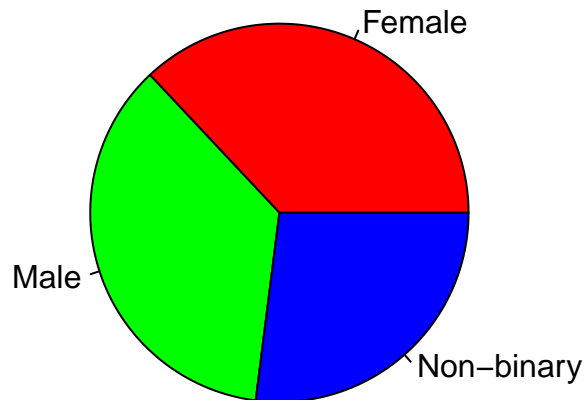
```
## [1] "Female"      "Male"        "Non-binary"
```

```
percentages_gender <- round(prop.table(table(data_transformed$Gender)) * 100, 2)
for (i in seq_along(genders)) {
  cat("Gender:", genders[i], "- Percentage:", percentages_gender[i], "%\n")
}
```

```
## Gender: Female - Percentage: 37 %
## Gender: Male - Percentage: 36 %
## Gender: Non-binary - Percentage: 27 %
```

```
pie(table(data_transformed$Gender), main = "Distribution per age",
    col = rainbow(length(unique(data_transformed$Gender))))
```

Distribution per age



```
labels = genders
```

3.1.2 Data distribution per platform

```
platforms <- unique(data_transformed$Platform)
print(platforms)

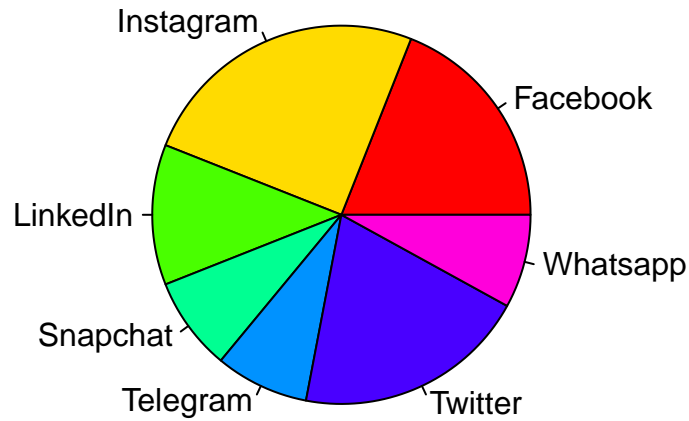
## [1] "Instagram" "Twitter"   "Facebook"  "LinkedIn"  "Whatsapp"  "Telegram"
## [7] "Snapchat"

percentages_platform <- round(prop.table(table(data_transformed$Platform)) * 100, 2)
for (i in seq_along(platforms)) {
  cat("Gender:", platforms[i], "- Percentage:", percentages_platform[i], "%\n")
}

## Gender: Instagram - Percentage: 19 %
## Gender: Twitter - Percentage: 25 %
## Gender: Facebook - Percentage: 12 %
## Gender: LinkedIn - Percentage: 8 %
## Gender: Whatsapp - Percentage: 8 %
## Gender: Telegram - Percentage: 20 %
## Gender: Snapchat - Percentage: 8 %

pie(table(data_transformed$Platform), main = "Distribution per platform",
    col = rainbow(length(unique(data_transformed$Platform))))
```

Distribution per platform



```
labels = platforms
```

3.1.3 Data distribution per age

```
ages <- unique(data_transformed$Age)
print(ages)

## [1] 25 30 22 28 33 21 27 24 29 31 23 26 34 35 32

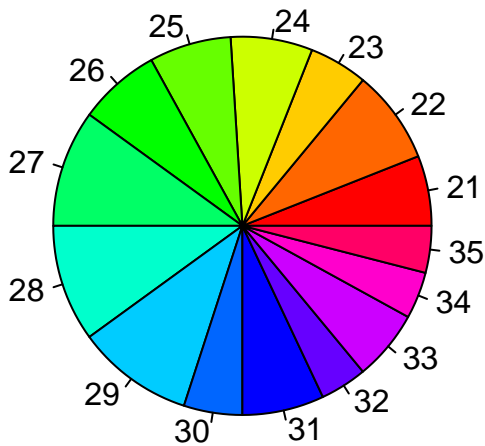
percentages_ages <- round(prop.table(table(data_transformed$Age)) * 100,
                           2)
for (i in seq_along(ages)) {
  cat("Gender:", ages[i], "- Percentage:", percentages_ages[i], "%\n")
}
```

```
## Gender: 25 - Percentage: 6 %
## Gender: 30 - Percentage: 8 %
## Gender: 22 - Percentage: 5 %
## Gender: 28 - Percentage: 7 %
## Gender: 33 - Percentage: 7 %
## Gender: 21 - Percentage: 7 %
## Gender: 27 - Percentage: 10 %
## Gender: 24 - Percentage: 10 %
## Gender: 29 - Percentage: 10 %
## Gender: 31 - Percentage: 5 %
## Gender: 23 - Percentage: 7 %
## Gender: 26 - Percentage: 4 %
```

```
## Gender: 34 - Percentage: 6 %
## Gender: 35 - Percentage: 4 %
## Gender: 32 - Percentage: 4 %

pie(table(data_transformed$Age), main = "Distribution per age",
    col = rainbow(length(unique(data_transformed$Age))))
```

Distribution per age



```
labels = ages
```

3.1.4 Data distribution per dominant emotion

```
emotions <- unique(data_transformed$Dominant_Emotion)
print(emotions)

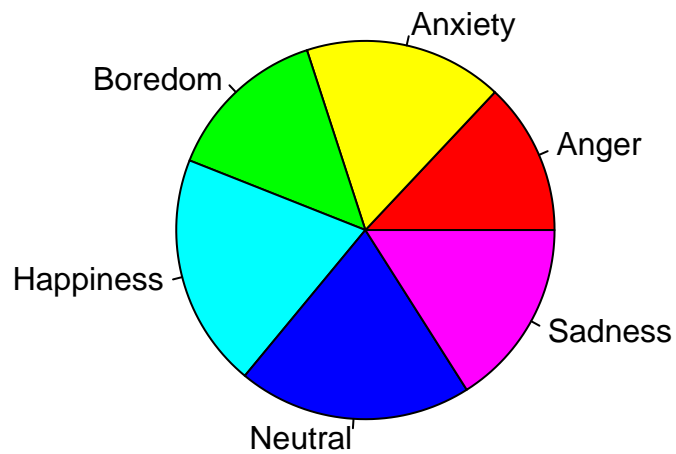
## [1] "Happiness" "Anger"      "Neutral"    "Anxiety"    "Boredom"    "Sadness"

percentages_emotions <- round(prop.table(table(data_transformed$Dominant_Emotion)) * 100, 2)
for (i in seq_along(emotions)) {
  cat("Gender:", emotions[i], "- Percentage:", percentages_emotions[i], "%\n")
}
```

```
## Gender: Happiness - Percentage: 13 %
## Gender: Anger - Percentage: 17 %
## Gender: Neutral - Percentage: 14 %
## Gender: Anxiety - Percentage: 20 %
## Gender: Boredom - Percentage: 20 %
## Gender: Sadness - Percentage: 16 %
```

```
pie(table(data_transformed$Dominant_Emotion), main = "Distribution per dominant emotion",  
    col = rainbow(length(unique(data_transformed$Dominant_Emotion))))
```

Distribution per dominant emotion

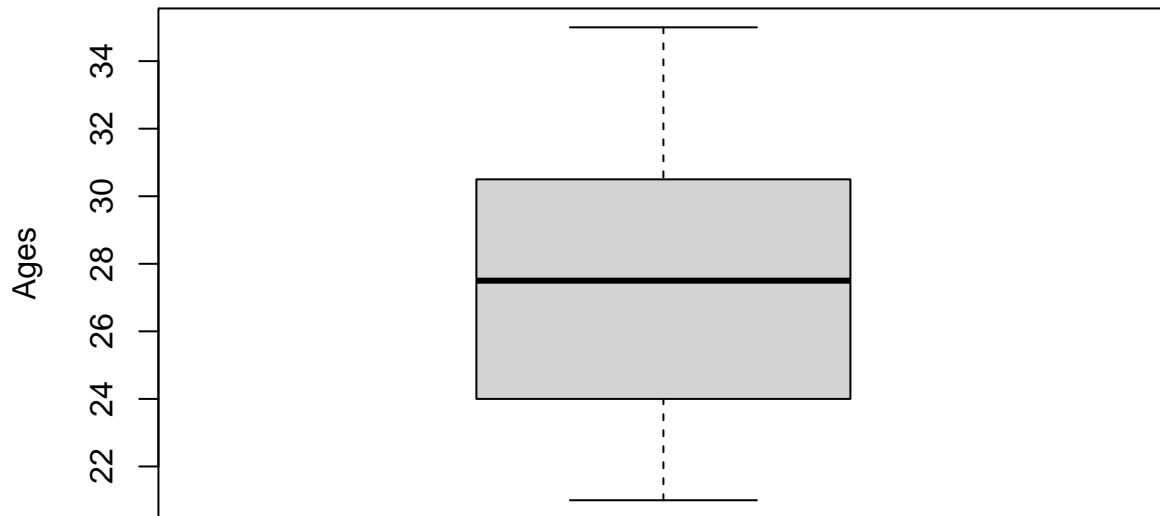


```
labels = emotions
```

3.2 Check if there are extreme ages

```
boxplot(data_transformed$Age, main = "Boxplot of ages",  
        ylab = "Ages")
```

Boxplot of ages



```
cat("Median: ",median(data_transformed$Age),"\n")
```

```
## Median: 27.5
```

```
cat("Quantiles 25%, 50%, 75%: ",quantile(data_transformed$Age, probs = c(0.25, 0.5, 0.75)))
```

```
## Quantiles 25%, 50%, 75%: 24 27.5 30.25
```

Median line -> Closer to Q3 (not too much) -> More people has less than 27 YO.

Box -> Q1 = 24 and Q3 = 30.25 ~ 30 -> 50% of the people has between 24 and 30 YO.

Whiskers -> Top whisker is longer than bottom whisker -> Ages above median are more dispersed

Outliers -> There are no outliers

3.3 Relation between used platform and dominant emotion

H0 -> There's no significant association between both variables

H1 -> There's significant association between both variables

```
chisq_data <- table(data_transformed$Platform, data_transformed$Dominant_Emotion)
chisq_data
```

```
##
##      Anger Anxiety Boredom Happiness Neutral Sadness
## Facebook      0      50      40         0       70      30
## Instagram     10      30       0       170      20      20
## LinkedIn       0      20      70         0       20      10
```

##	Snapchat	0	20	0	10	20	30
##	Telegram	10	10	10	0	30	20
##	Twitter	80	20	20	10	20	50
##	Whatsapp	30	20	0	10	20	0

```
# Alpha = 0.05 -> CL = 95%
chisq.test(chisq_data, correct = FALSE)
```

```
##
## Pearson's Chi-squared test
##
## data:  chisq_data
## X-squared = 1003.9, df = 30, p-value < 2.2e-16
```

The p value < 0.05(alpha) -> There's enough evidence to refuse H0 with a 95% confidence level

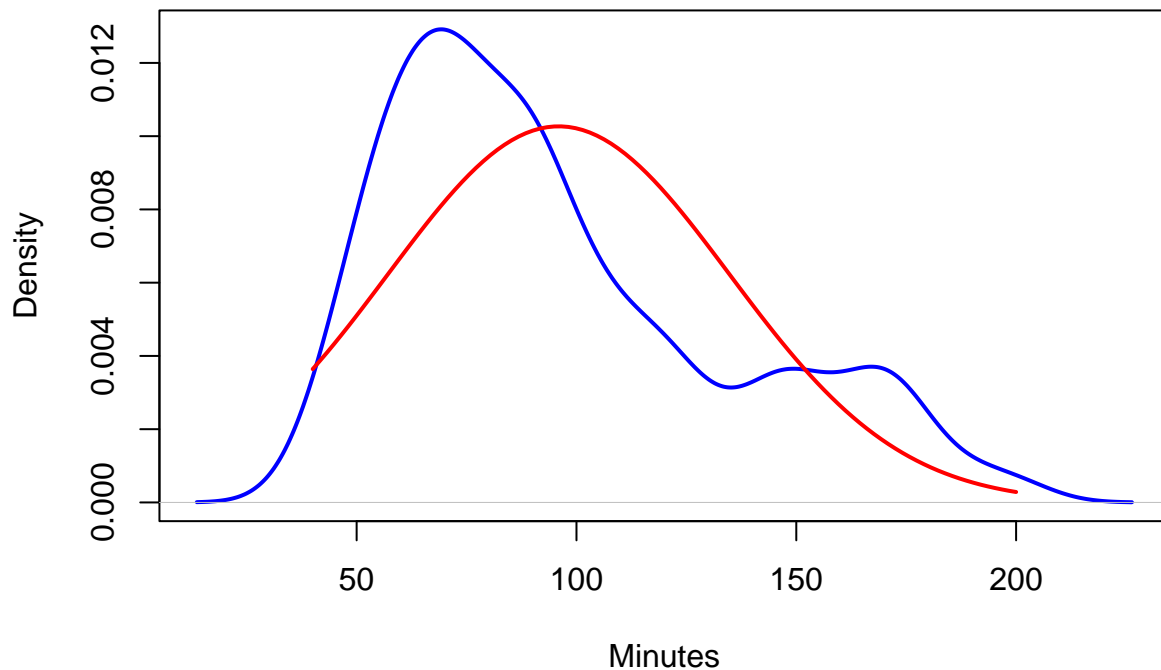
3.4 Dominant emotion depending on the minutes per day spend on social media

H0 -> The spent time on social media is the same regardless of the dominant emotion

H1 -> At least one emotion spends more/less time on social media

```
plot(density(data_transformed$Minutes_Per_Day),
     main = "Minutes dedicated to social media per day",
     xlab = "Minutes",
     ylab = "Density",
     col = "blue",
     lwd = 2)
values_normal_distribution <- seq(min(data_transformed$Minutes_Per_Day),
                                max(data_transformed$Minutes_Per_Day),
                                length = 100)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Minutes_Per_Day)),
                             sd = sd(data_transformed$Minutes_Per_Day))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)
```


Minutes dedicated to social media per day



Variable Minutes_Per_Day -> No normal distribution -> ANOVA no possible

```
kruskal <- kruskal.test(Minutes_Per_Day ~ Dominant_Emotion, data = data_transformed)
kruskal
```

```
##
## Kruskal-Wallis rank sum test
##
## data: Minutes_Per_Day by Dominant_Emotion
## Kruskal-Wallis chi-squared = 475.45, df = 5, p-value < 2.2e-16
```

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation

```
library(dunn.test)
```

```
results_dunn <- dunn.test(data_transformed$Minutes_Per_Day,
                          data_transformed$Dominant_Emotion, method = "bonferroni")
```

```
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 475.4502, df = 5, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | Anger Anxiety Boredom Happiness Neutral
## -----+-----
```

```

## Anxiety | 0.609369
##         | 1.0000
##         |
## Boredom | 8.552439 8.505242
##         | 0.0000* 0.0000*
##         |
## Happiness | -10.06400 -11.54933 -19.74230
##         | 0.0000* 0.0000* 0.0000*
##         |
## Neutral | 4.740492 4.438969 -4.606576 16.67876
##         | 0.0000* 0.0001* 0.0000* 0.0000*
##         |
## Sadness | 2.720632 2.271944 -6.225303 13.71841 -2.006476
##         | 0.0489 0.1732 0.0000* 0.0000* 0.3360
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
print(results_dunn)

## $chi2
## [1] 475.4502
##
## $Z
## [1] 0.6093694 8.5524400 8.5052428 -10.0640024 -11.5493378 -19.7423090
## [7] 4.7404926 4.4389697 -4.6065762 16.6787652 2.7206324 2.2719444
## [13] -6.2253040 13.7184142 -2.0064764
##
## $P
## [1] 2.711398e-01 6.025671e-18 9.060798e-18 3.984555e-24 3.719570e-31
## [6] 4.670510e-87 1.065996e-06 4.519528e-06 2.046765e-06 9.352266e-63
## [11] 3.257859e-03 1.154493e-02 2.403117e-10 3.938471e-43 2.240272e-02
##
## $P.adjusted
## [1] 1.000000e+00 9.038506e-17 1.359120e-16 5.976832e-23 5.579356e-30
## [6] 7.005765e-86 1.598994e-05 6.779291e-05 3.070148e-05 1.402840e-61
## [11] 4.886788e-02 1.731740e-01 3.604675e-09 5.907706e-42 3.360408e-01
##
## $comparisons
## [1] "Anger - Anxiety" "Anger - Boredom" "Anxiety - Boredom"
## [4] "Anger - Happiness" "Anxiety - Happiness" "Boredom - Happiness"
## [7] "Anger - Neutral" "Anxiety - Neutral" "Boredom - Neutral"
## [10] "Happiness - Neutral" "Anger - Sadness" "Anxiety - Sadness"
## [13] "Boredom - Sadness" "Happiness - Sadness" "Neutral - Sadness"

```

Significant difference = Between both emotions, one of them spends more/less time on social media than the other emotion.

Considering significance = 0.05 -> CL = 95%: - "Anger - Boredom"

- "Anxiety - Boredom"
- "Anger - Happiness"
- "Anxiety - Happiness"
- "Boredom - Happiness"

- “Anger - Neutral”
- “Anxiety - Neutral”
- “Boredom - Neutral”
- “Happiness - Neutral”
- “Anger - Sadness”

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Minutes_Per_Day)) +
  geom_violin() +
  scale_y_continuous(limits = c(0, 200),
                     breaks = seq(0, 200, by = 20)) +
  labs(title = "Distribution of the minutes dedicated to social media
              per dominant emotion",
       x = "Dominant emotion",
       y = "Minutes per day")
```

Distribution of the minutes dedicated to social media per dominant emotion



Taking a look at the graphic above, H1 is confirmed.

```
minutes_platform <- data_transformed %>%
  group_by(Platform) %>%
  summarise(suma = sum(Minutes_Per_Day)) %>%
  arrange(desc(suma))
print(minutes_platform)
```

```
## # A tibble: 7 x 2
##   Platform  suma
##   <chr>    <int>
## 1 Instagram 38350
## 2 Twitter  16750
## 3 Facebook 13700
## 4 Snapchat  7200
## 5 Whatsapp  7000
## 6 LinkedIn  6700
## 7 Telegram  6250
```

3.5 Dominant emotion depending on the posts per day spend on social media

H0 -> The daily posts are the same regardless of the dominant emotion

H1 -> At least one emotion has more/less daily posts

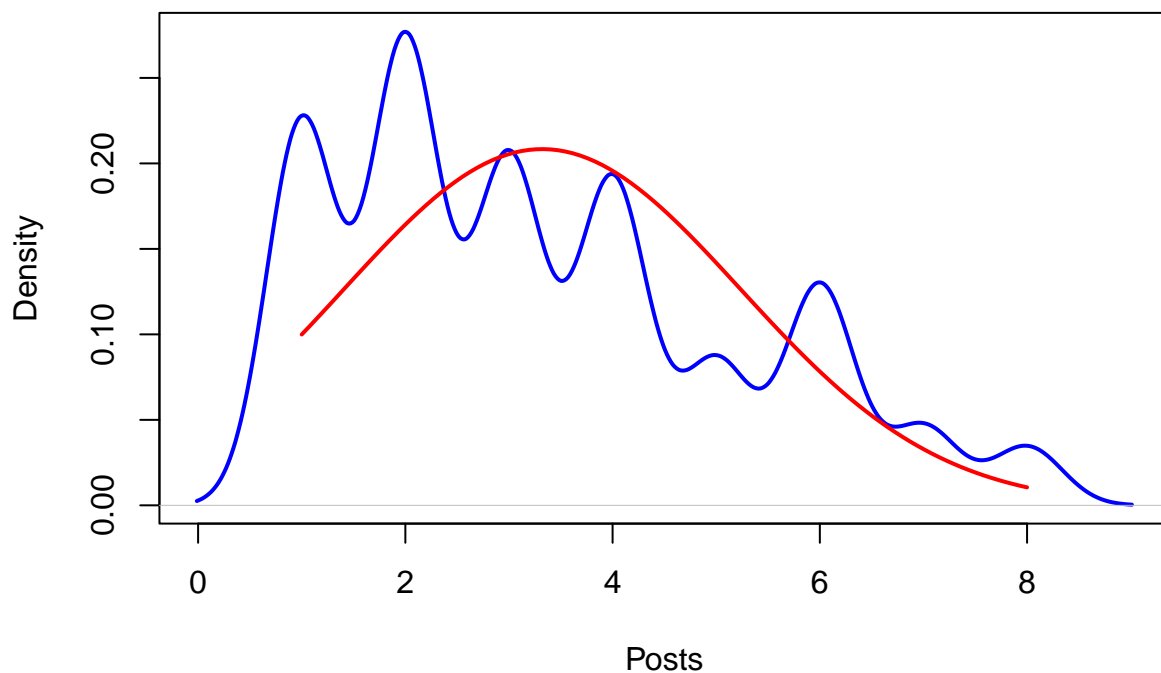
```
plot(density(data_transformed$Posts_Per_Day),
     main = "Posts per day on social media per day",
     xlab = "Posts",
```

```

ylab = "Density",
col = "blue",
lwd = 2)
values_normal_distribution <- seq(min(data_transformed$Posts_Per_Day),
                                max(data_transformed$Posts_Per_Day),
                                length = 100)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Posts_Per_Day)),
                             sd = sd(data_transformed$Posts_Per_Day))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)

```

Posts per day on social media per day



Variable Posts_Per_Day -> No normal distribution -> ANOVA no possible

```

kruskal <- kruskal.test(Posts_Per_Day ~ Dominant_Emotion, data = data_transformed)
kruskal

```

```

##
## Kruskal-Wallis rank sum test
##
## data: Posts_Per_Day by Dominant_Emotion
## Kruskal-Wallis chi-squared = 474.09, df = 5, p-value < 2.2e-16

```

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation

```

results_dunn <- dunn.test(data_transformed$Posts_Per_Day,
                          data_transformed$Dominant_Emotion, method = "bonferroni")

```

```

## Kruskal-Wallis rank sum test

```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 474.0937, df = 5, p-value = 0
##
##
##              Comparison of x by group
##              (Bonferroni)
## Col Mean-|
## Row Mean |      Anger      Anxiety      Boredom      Happines      Neutral
## -----+-----
## Anxiety |      4.530421
##          |      0.0000*
##          |
## Boredom |      11.31303      7.448500
##          |      0.0000*      0.0000*
##          |
## Happines |      -7.023301     -12.64480     -19.68489
##          |      0.0000*      0.0000*      0.0000*
##          |
## Neutral |       7.200801      2.716701     -5.142541      16.02489
##          |      0.0000*      0.0495      0.0000*      0.0000*
##          |
## Sadness |       3.894544     -0.617199     -7.932991      11.79552     -3.312889
##          |      0.0007*      1.0000      0.0000*      0.0000*      0.0069*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
print(results_dunn)

## $chi2
## [1] 474.0937
##
## $Z
## [1]  4.5304217 11.3130314  7.4485001 -7.0233013 -12.6448033 -19.6848957
## [7]  7.2008010  2.7167015 -5.1425414 16.0248939  3.8945449 -0.6171998
## [13] -7.9329919 11.7955258 -3.3128891
##
## $P
## [1] 2.943303e-06 5.655637e-30 4.720368e-14 1.083431e-12 5.976165e-37
## [6] 1.452671e-86 2.992992e-13 3.296802e-03 1.355233e-07 4.281844e-58
## [11] 4.919163e-05 2.685515e-01 1.069642e-15 2.058108e-32 4.616879e-04
##
## $P.adjusted
## [1] 4.414955e-05 8.483455e-29 7.080553e-13 1.625147e-11 8.964247e-36
## [6] 2.179007e-85 4.489488e-12 4.945204e-02 2.032850e-06 6.422766e-57
## [11] 7.378745e-04 1.000000e+00 1.604463e-14 3.087163e-31 6.925319e-03
##
## $comparisons
## [1] "Anger - Anxiety"      "Anger - Boredom"      "Anxiety - Boredom"
## [4] "Anger - Happiness"    "Anxiety - Happiness"  "Boredom - Happiness"
## [7] "Anger - Neutral"      "Anxiety - Neutral"    "Boredom - Neutral"
## [10] "Happiness - Neutral"   "Anger - Sadness"      "Anxiety - Sadness"
## [13] "Boredom - Sadness"    "Happiness - Sadness"  "Neutral - Sadness"
```

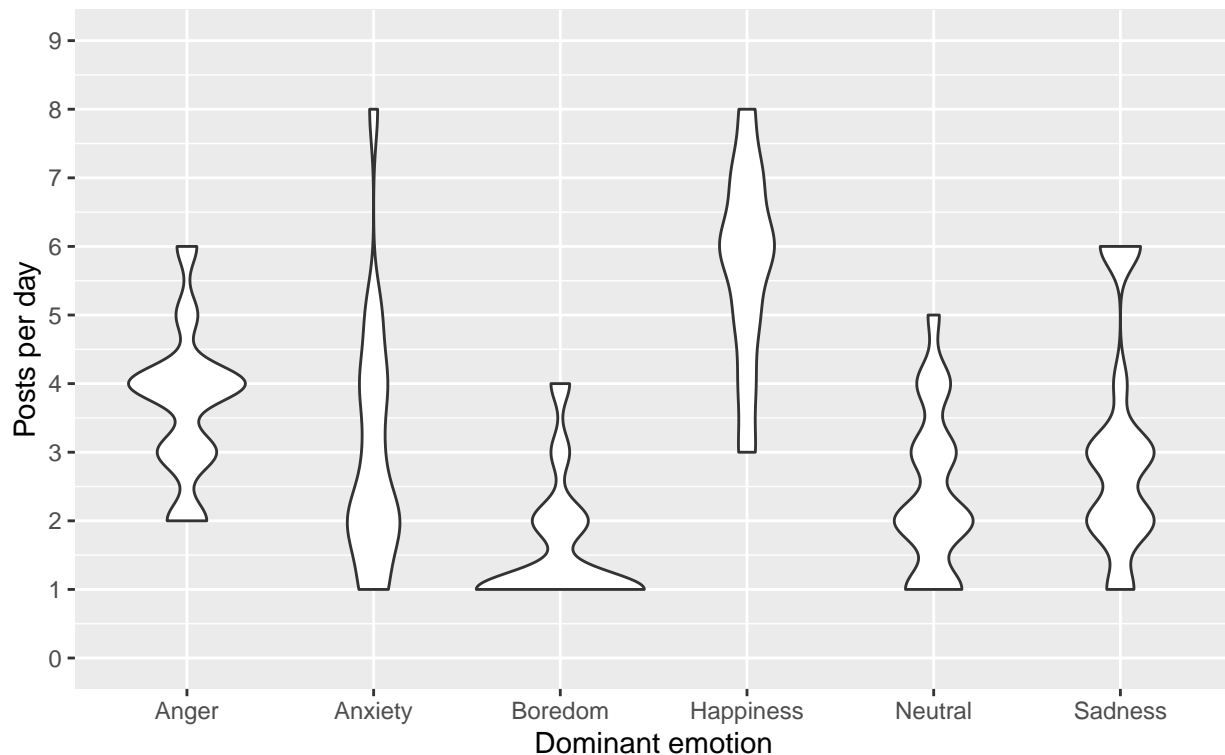
Significant difference = Between both emotions, one of them posts more/less on social media than the other emotion.

Considering significance = 0.05 -> CL = 95%: - “Anger - Anxiety”

- “Anger - Boredom”
- “Anxiety - Boredom”
- “Anger - Happiness”
- “Anxiety - Happiness”
- “Boredom - Happiness”
- “Anger - Neutral”
- “Anxiety - Neutral”
- “Boredom - Neutral”
- “Happiness - Neutral”
- “Anger - Sadness”
- “Boredom - Sadness”
- “Happiness - Sadness”
- “Neutral - Sadness”

```
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Posts_Per_Day)) +  
  geom_violin() +  
  scale_y_continuous(limits = c(0, 9),  
                     breaks = seq(0, 9, by = 1)) +  
  labs(  
    title = "Distribution of the posts per day on social media  
per dominant emotion",  
    x = "Dominant emotion",  
    y = "Posts per day")
```

Distribution of the posts per day on social media per dominant emotion



Taking a look at the graphic above, H1 is confirmed.

```
posts_platform <- data_transformed %>%
  group_by(Platform) %>%
  summarise(suma = sum(Posts_Per_Day)) %>%
  arrange(desc(suma))
print(posts_platform)
```

```
## # A tibble: 7 x 2
##   Platform   suma
##   <chr>     <int>
## 1 Instagram 1450
## 2 Twitter   681
## 3 Facebook  370
## 4 Whatsapp  240
## 5 Telegram  220
## 6 Snapchat  210
## 7 LinkedIn  150
```

3.6 Dominant emotion depending on the likes received per day spend on social media

H0 -> The likes received are the same regardless of the dominant emotion

H1 -> At least one emotion has more/less likes

```
plot(density(data_transformed$Likes_Received_Per_Day),
     main = "Likes per day on social media per day",
```

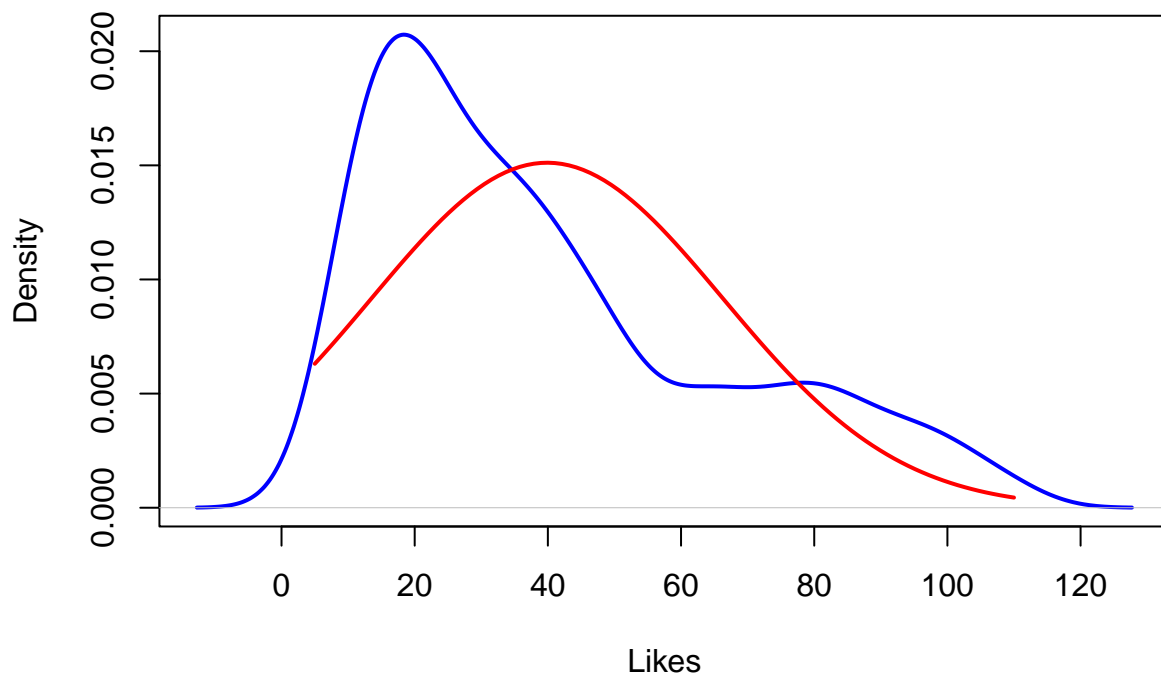


```

xlab = "Likes",
ylab = "Density",
col = "blue",
lwd = 2)
values_normal_distribution <- seq(min(data_transformed$Likes_Received_Per_Day),
                                max(data_transformed$Likes_Received_Per_Day),
                                length = 100)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Likes_Received_Per_Day)),
                             sd = sd(data_transformed$Likes_Received_Per_Day))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)

```

Likes per day on social media per day



Variable Likes_Received_Per_Day -> No normal distribution -> ANOVA no possible

```

kruskal <- kruskal.test(Likes_Received_Per_Day ~ Dominant_Emotion, data = data_transformed)
kruskal

```

```

##
## Kruskal-Wallis rank sum test
##
## data: Likes_Received_Per_Day by Dominant_Emotion
## Kruskal-Wallis chi-squared = 529.6, df = 5, p-value < 2.2e-16

```

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation

```

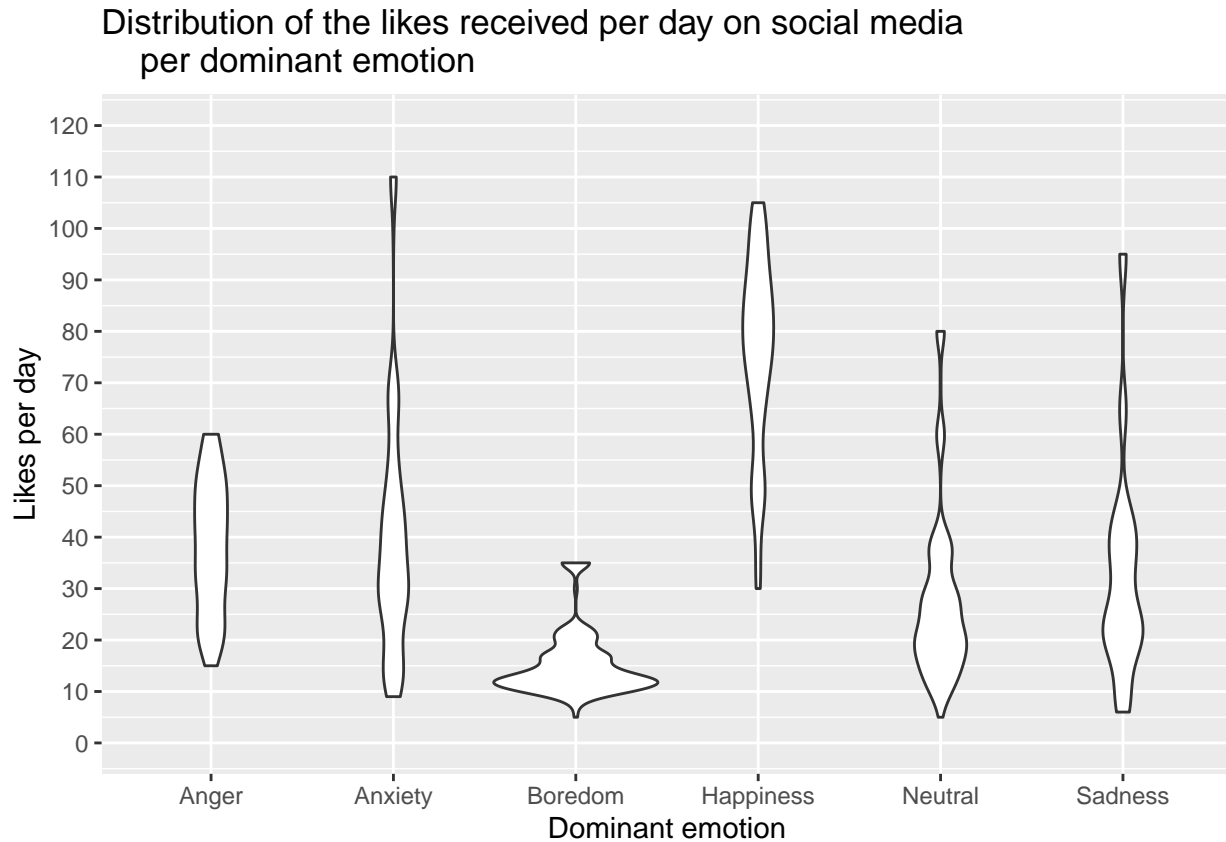
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Likes_Received_Per_Day)) +
  geom_violin() +
  scale_y_continuous(limits = c(0, 120),

```

```

breaks = seq(0, 120, by = 10)) +
labs(
  title = "Distribution of the likes received per day on social media
per dominant emotion",
  x = "Dominant emotion",
  y = "Likes per day")

```



Taking a look at the graphic above, H1 is confirmed.

```

likes_platform <- data_transformed %>%
  group_by(Platform) %>%
  summarise(suma = sum(Likes_Received_Per_Day)) %>%
  arrange(desc(suma))
print(likes_platform)

```

```

## # A tibble: 7 x 2
##   Platform  suma
##   <chr>    <int>
## 1 Instagram 19818
## 2 Twitter   7049
## 3 Facebook  3748
## 4 Whatsapp  2916
## 5 Snapchat  2436
## 6 Telegram  2386
## 7 LinkedIn  1545

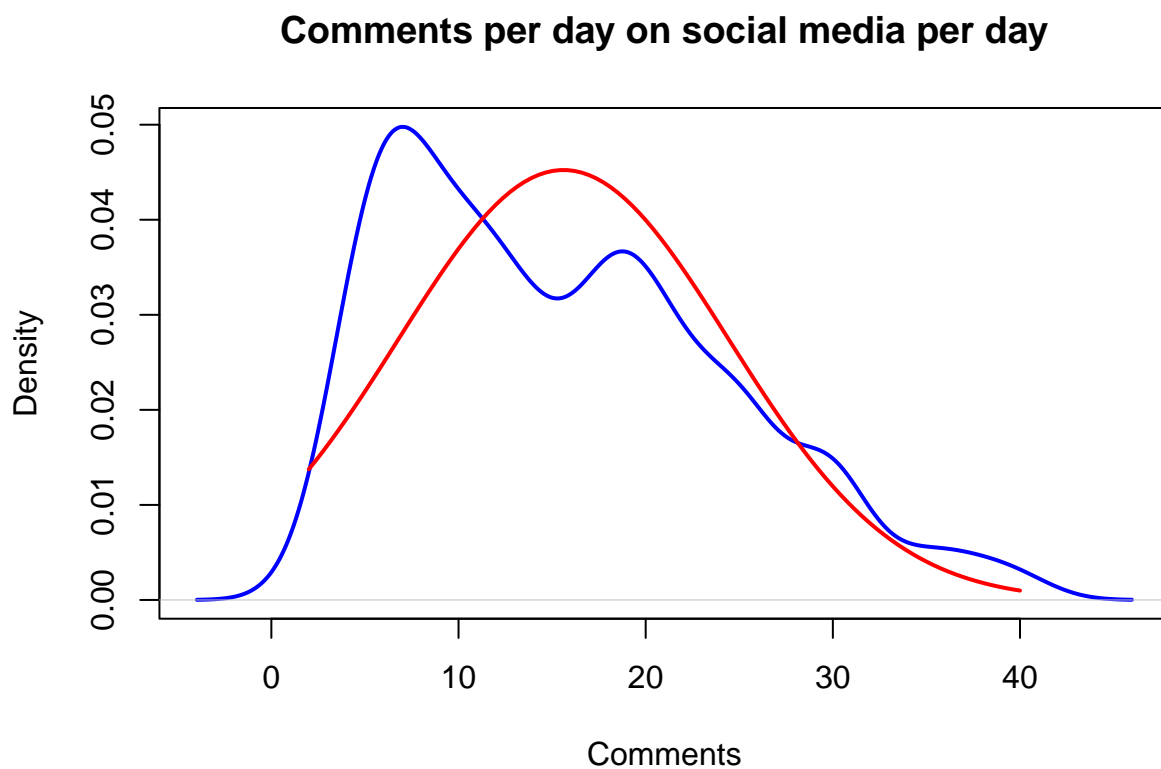
```

3.7 Dominant emotion depending on the comments received per day spend on social media

H0 -> The comments received are the same regardless of the dominant emotion

H1 -> At least one emotion has more/less comments

```
plot(density(data_transformed$Comments_Received_Per_Day),
     main = "Comments per day on social media per day",
     xlab = "Comments",
     ylab = "Density",
     col = "blue",
     lwd = 2)
values_normal_distribution <- seq(min(data_transformed$Comments_Received_Per_Day),
                                max(data_transformed$Comments_Received_Per_Day),
                                length = 100)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Comments_Received_Per_Day)),
                             sd = sd(data_transformed$Comments_Received_Per_Day))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)
```



Variable Comments_Received_Per_Day -> No normal distribution -> ANOVA no possible

```
kruskal <- kruskal.test(Comments_Received_Per_Day ~ Dominant_Emotion, data = data_transformed)
kruskal
```

```
##
## Kruskal-Wallis rank sum test
##
```

```
## data: Comments_Received_Per_Day by Dominant_Emotion
## Kruskal-Wallis chi-squared = 553.55, df = 5, p-value < 2.2e-16

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation
```

```
results_dunn <- dunn.test(data_transformed$Comments_Received_Per_Day,
                          data_transformed$Dominant_Emotion, method = "bonferroni")
```

```
## Kruskal-Wallis rank sum test
```

```
##
```

```
## data: x and group
```

```
## Kruskal-Wallis chi-squared = 553.5507, df = 5, p-value = 0
```

```
##
```

```
##
```

```
## Comparison of x by group
## (Bonferroni)
```

```
## Col Mean-|
## Row Mean |      Anger      Anxiety      Boredom      Happines      Neutral
## -----+-----
```

	Anger	Anxiety	Boredom	Happines	Neutral
Anxiety	3.834512				
	0.0009*				
Boredom	12.50938	9.435710			
	0.0000*	0.0000*			
Happines	-6.936749	-11.77408	-20.91875		
	0.0000*	0.0000*	0.0000*		
Neutral	8.578942	4.982284	-5.055914	17.48000	
	0.0000*	0.0000*	0.0000*	0.0000*	
Sadness	5.857789	2.223511	-7.189008	13.88916	-2.591136
	0.0000*	0.1964	0.0000*	0.0000*	0.0717

```
##
```

```
## alpha = 0.05
```

```
## Reject Ho if p <= alpha/2
```

```
print(results_dunn)
```

```
## $chi2
```

```
## [1] 553.5507
```

```
##
```

```
## $Z
```

```
## [1] 3.834513 12.509388 9.435710 -6.936749 -11.774090 -20.918757
```

```
## [7] 8.578943 4.982284 -5.055914 17.480001 5.857789 2.223511
```

```
## [13] -7.189009 13.889167 -2.591136
```

```
##
```

```
## $P
```

```
## [1] 6.290670e-05 3.316685e-36 1.941773e-21 2.006123e-12 2.654348e-32
```

```
## [6] 1.807069e-97 4.787452e-18 3.141901e-07 2.141669e-07 1.017478e-68
```

```
## [11] 2.345344e-09 1.309067e-02 3.263166e-13 3.684376e-44 4.782981e-03
```

```
##
```

```
## $P.adjusted
```

```
## [1] 9.436006e-04 4.975028e-35 2.912660e-20 3.009185e-11 3.981523e-31
```

```
## [6] 2.710603e-96 7.181178e-17 4.712852e-06 3.212503e-06 1.526217e-67
```

```
## [11] 3.518016e-08 1.963601e-01 4.894750e-12 5.526564e-43 7.174472e-02
```

```
##
## $comparisons
## [1] "Anger - Anxiety"      "Anger - Boredom"      "Anxiety - Boredom"
## [4] "Anger - Happiness"    "Anxiety - Happiness"   "Boredom - Happiness"
## [7] "Anger - Neutral"      "Anxiety - Neutral"     "Boredom - Neutral"
## [10] "Happiness - Neutral"   "Anger - Sadness"       "Anxiety - Sadness"
## [13] "Boredom - Sadness"    "Happiness - Sadness"   "Neutral - Sadness"
```

Significant difference = Between both emotions, one of them posts more/less on social media than the other emotion.

Considering significance = 0.05 -> CL = 95%:

- “Anger - Anxiety”
- “Anger - Boredom”
- “Anxiety - Boredom”
- “Anger - Happiness”
- “Anxiety - Happiness”
- “Boredom - Happiness”
- “Anger - Neutral”
- “Anxiety - Neutral”
- “Boredom - Neutral”
- “Happiness - Neutral”
- “Anger - Sadness”
- “Anxiety - Sadness”
- “Happiness - Sadness”
- “Neutral - Sadness”

```
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Comments_Received_Per_Day)) +
  geom_violin() +
  scale_y_continuous(limits = c(0, 45),
                     breaks = seq(0, 45, by = 10)) +
  labs(
    title = "Distribution of the comments received per day on social media per
    dominant emotion",
    x = "Dominant emotion",
    y = "Comments per day")
```

Distribution of the comments received per day on social media per dominant emotion



Taking a look at the graphic above, H1 is confirmed.

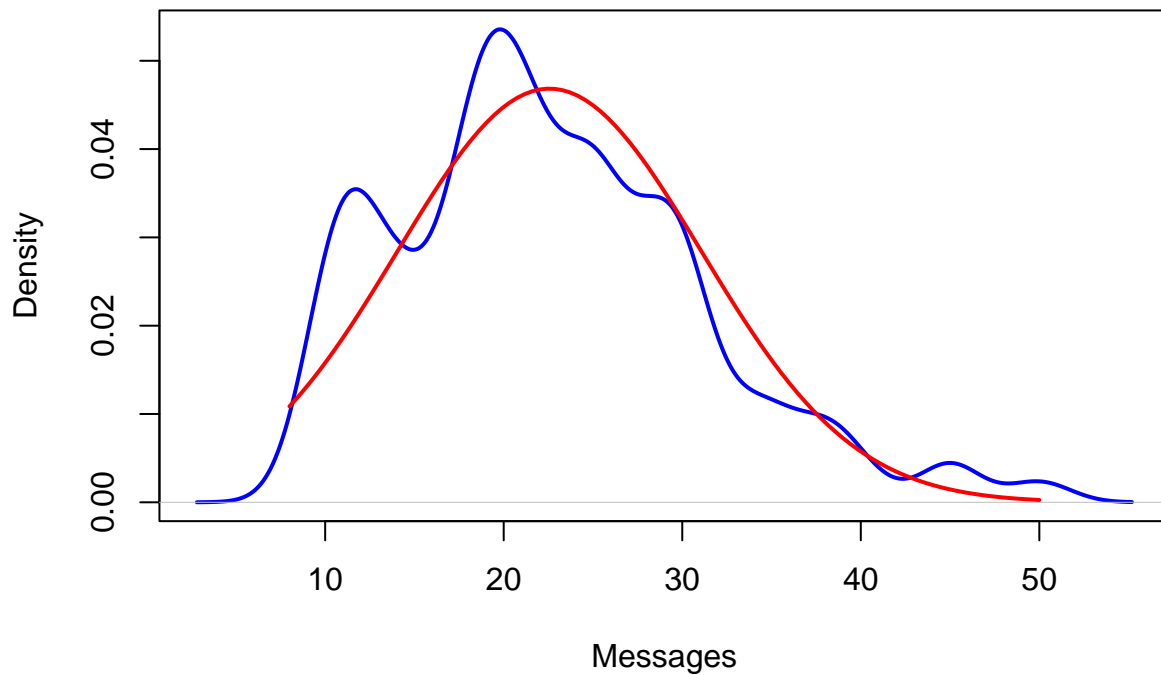
3.8 Dominant emotion depending on the messages sent per day spend on social media

H0 -> The messages sent are the same regardless of the dominant emotion

H1 -> At least one emotion has more/less messages

```
plot(density(data_transformed$Messages_Sent_Per_Day),
     main = "Messages per day on social media per day",
     xlab = "Messages",
     ylab = "Density",
     col = "blue",
     lwd = 2)
values_normal_distribution <- seq(min(data_transformed$Messages_Sent_Per_Day),
                                max(data_transformed$Messages_Sent_Per_Day),
                                length = 100)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Messages_Sent_Per_Day)),
                             sd = sd(data_transformed$Messages_Sent_Per_Day))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)
```

Messages per day on social media per day



Variable Comments_Received_Per_Day -> No normal distribution -> ANOVA no possible

```
kruskal <- kruskal.test(Messages_Sent_Per_Day ~ Dominant_Emotion, data = data_transformed)
kruskal
```

```
##
## Kruskal-Wallis rank sum test
##
## data: Messages_Sent_Per_Day by Dominant_Emotion
## Kruskal-Wallis chi-squared = 465.51, df = 5, p-value < 2.2e-16
```

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation

```
messages_dunn <- dunn.test(data_transformed$Messages_Sent_Per_Day,
                           data_transformed$Dominant_Emotion, method = "bonferroni")
```

```
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 465.5125, df = 5, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | Anger Anxiety Boredom Happines Neutral
## -----+-----
## Anxiety | -1.050281
```

```

##          |      1.0000
##          |
## Boredom  |    9.423240    11.12887
##          |    0.0000*    0.0000*
##          |
## Happines |   -9.329874   -8.902889   -19.95426
##          |    0.0000*    0.0000*    0.0000*
##          |
## Neutral  |    4.650077    6.194941   -5.661522    15.74983
##          |    0.0000*    0.0000*    0.0000*    0.0000*
##          |
## Sadness  |    1.826229    3.068676   -8.054360    11.94295   -2.906131
##          |    0.5086    0.0161*    0.0000*    0.0000*    0.0274
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
print(results_dunn)

## $chi2
## [1] 553.5507
##
## $Z
## [1]  3.834513 12.509388  9.435710 -6.936749 -11.774090 -20.918757
## [7]  8.578943  4.982284 -5.055914 17.480001  5.857789  2.223511
## [13] -7.189009 13.889167 -2.591136
##
## $P
## [1] 6.290670e-05 3.316685e-36 1.941773e-21 2.006123e-12 2.654348e-32
## [6] 1.807069e-97 4.787452e-18 3.141901e-07 2.141669e-07 1.017478e-68
## [11] 2.345344e-09 1.309067e-02 3.263166e-13 3.684376e-44 4.782981e-03
##
## $P.adjusted
## [1] 9.436006e-04 4.975028e-35 2.912660e-20 3.009185e-11 3.981523e-31
## [6] 2.710603e-96 7.181178e-17 4.712852e-06 3.212503e-06 1.526217e-67
## [11] 3.518016e-08 1.963601e-01 4.894750e-12 5.526564e-43 7.174472e-02
##
## $comparisons
## [1] "Anger - Anxiety"      "Anger - Boredom"      "Anxiety - Boredom"
## [4] "Anger - Happiness"    "Anxiety - Happiness"  "Boredom - Happiness"
## [7] "Anger - Neutral"      "Anxiety - Neutral"    "Boredom - Neutral"
## [10] "Happiness - Neutral"  "Anger - Sadness"      "Anxiety - Sadness"
## [13] "Boredom - Sadness"    "Happiness - Sadness"  "Neutral - Sadness"

```

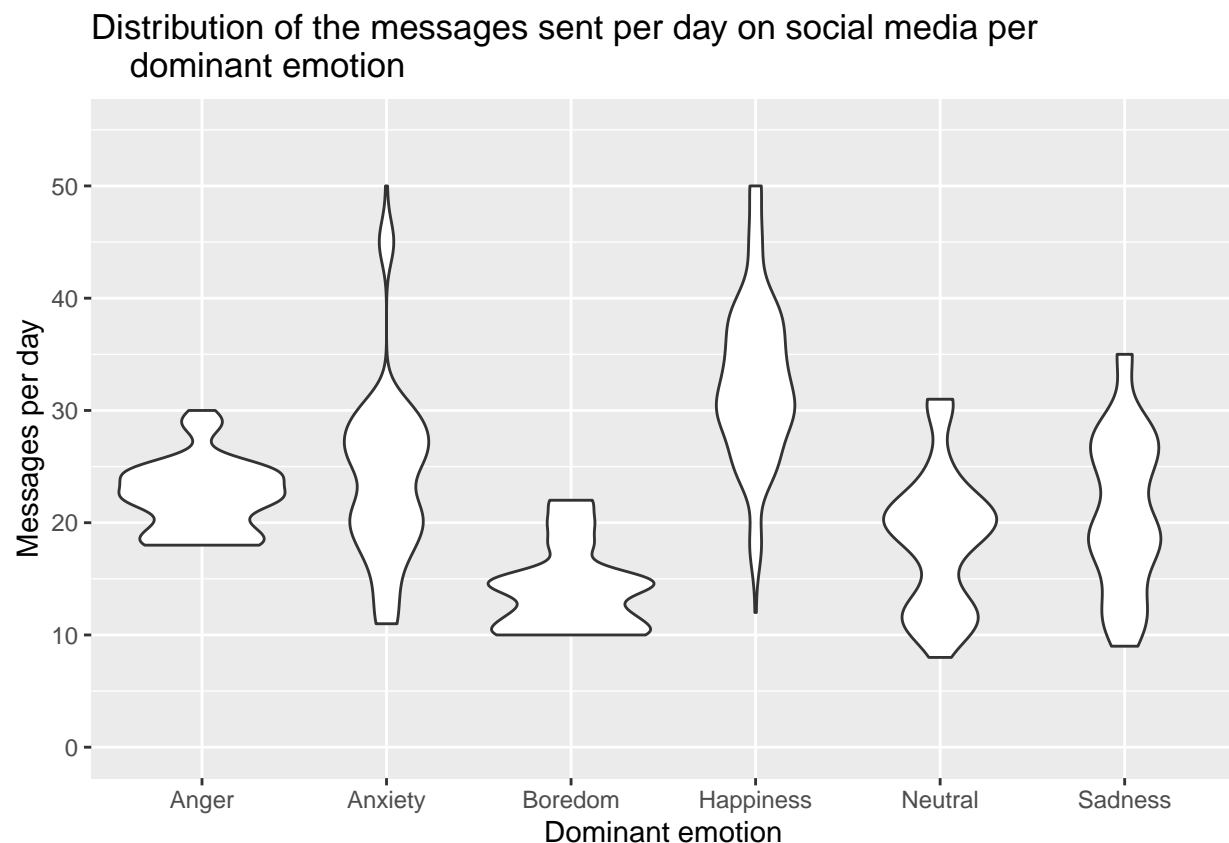
Significant difference = Between both emotions, one of them send more/less messages on social media than the other emotion.

Considering significance = 0.05 -> CL = 95%:

- "Anger - Anxiety"
- "Anger - Boredom"
- "Anxiety - Boredom"
- "Anger - Happiness"
- "Anxiety - Happiness"

- “Boredom - Happiness”
- “Anger - Neutral”
- “Anxiety - Neutral”
- “Boredom - Neutral”
- “Happiness - Neutral”
- “Anger - Sadness”
- “Anxiety - Sadness”
- “Boredom - Sadness”

```
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Messages_Sent_Per_Day)) +
  geom_violin() +
  scale_y_continuous(limits = c(0, 55),
    breaks = seq(0, 55, by = 10)) +
  labs(
    title = "Distribution of the messages sent per day on social media per
    dominant emotion",
    x = "Dominant emotion",
    y = "Messages per day")
```



Taking a look at the graphic above, H1 is confirmed.

3.9 Relation between dominant emotion and gender

H0 -> There's no significant association between both variables

H1 -> There's significant association between both variables

```
chisq_data_2 <- table(data_transformed$Gender, data_transformed$Dominant_Emotion)
chisq_data_2
```

```
##
##           Anger Anxiety Boredom Happiness Neutral Sadness
## Female      60      60      30      110      60      50
## Male        60      60      60       70      50      60
## Non-binary  10      50      50       20      90      50
```

```
# Alpha = 0.05 -> CL = 95%
chisq.test(chisq_data_2, correct = FALSE)
```

```
##
## Pearson's Chi-squared test
##
## data:  chisq_data_2
## X-squared = 115.03, df = 10, p-value < 2.2e-16
```

The p value < 0.05(alpha) -> There's enough evidence to refuse H0 with a 95% confidence level

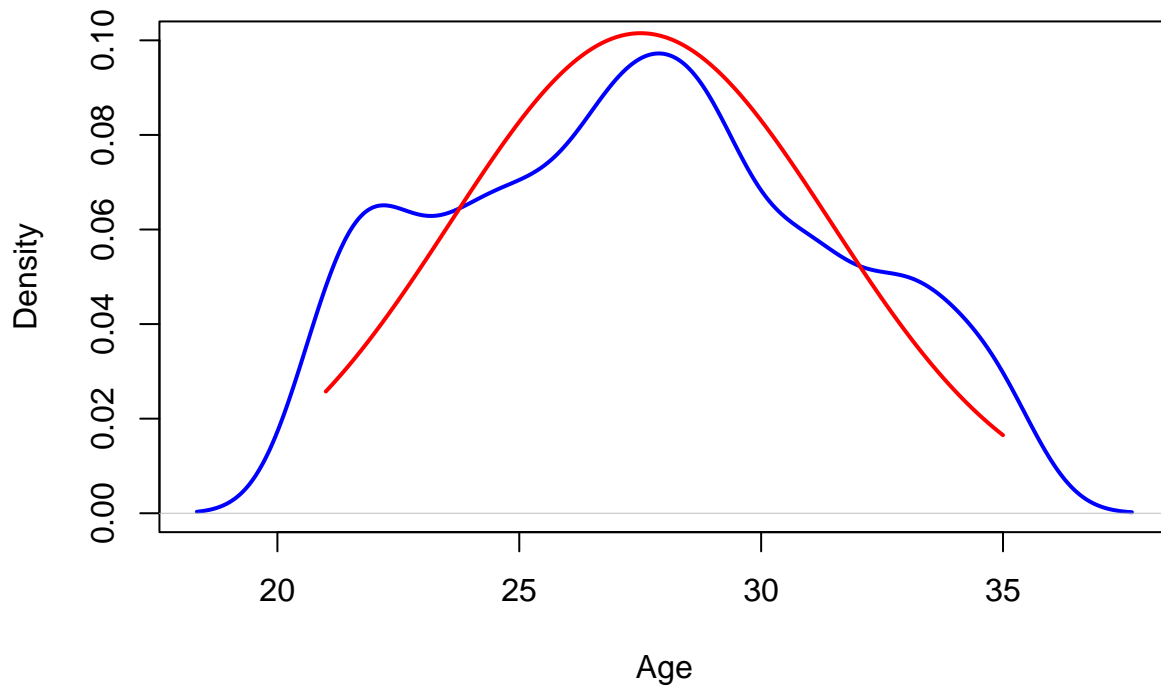
3.10 Dominant emotion depending on age

H0 -> The age is the same regardless of the dominant emotion

H1 -> At least one emotion has more/less age

```
plot(density(data_transformed$Age),
     main = "Age on social media per day",
     xlab = "Age",
     ylab = "Density",
     col = "blue",
     lwd = 2,
     ylim = c(0, 0.10))
values_normal_distribution <- seq(min(data_transformed$Age),
                                max(data_transformed$Age),
                                length = 1000)
normal_distribution <- dnorm(values_normal_distribution,
                             mean = abs(mean(data_transformed$Age)),
                             sd = sd(data_transformed$Age))
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)
```

Age on social media per day



Variable Age -> No normal distribution -> ANOVA no possible

```
kruskal <- kruskal.test(Age ~ Dominant_Emotion, data = data_transformed)
kruskal
```

```
##
## Kruskal-Wallis rank sum test
##
## data: Age by Dominant_Emotion
## Kruskal-Wallis chi-squared = 123.95, df = 5, p-value < 2.2e-16
```

p < 0.05 -> There's enough evidence to refuse H0 with a 95% confidence level -> Suggests relation

```
messages_dunn <- dunn.test(data_transformed$Age,
                           data_transformed$Dominant_Emotion, method = "bonferroni")
```

```
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 123.9519, df = 5, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean |      Anger      Anxiety      Boredom      Happines      Neutral
## -----+-----
## Anxiety |      0.856336
```

```

##          |      1.0000
##          |
## Boredom  | -3.757854 -4.884667
##          |  0.0013*  0.0000*
##          |
## Happiness |  0.704750 -0.195312  4.874129
##          |  1.0000  1.0000  0.0000*
##          |
## Neutral  |  6.067731  5.596508 10.35710  6.041942
##          |  0.0000*  0.0000*  0.0000*  0.0000*
##          |
## Sadness  |  3.535984  2.884767  7.562816  3.187854 -2.508542
##          |  0.0030*  0.0294  0.0000*  0.0107*  0.0909
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
print(results_dunn)

## $chi2
## [1] 553.5507
##
## $Z
## [1]  3.834513 12.509388  9.435710 -6.936749 -11.774090 -20.918757
## [7]  8.578943  4.982284 -5.055914 17.480001  5.857789  2.223511
## [13] -7.189009 13.889167 -2.591136
##
## $P
## [1] 6.290670e-05 3.316685e-36 1.941773e-21 2.006123e-12 2.654348e-32
## [6] 1.807069e-97 4.787452e-18 3.141901e-07 2.141669e-07 1.017478e-68
## [11] 2.345344e-09 1.309067e-02 3.263166e-13 3.684376e-44 4.782981e-03
##
## $P.adjusted
## [1] 9.436006e-04 4.975028e-35 2.912660e-20 3.009185e-11 3.981523e-31
## [6] 2.710603e-96 7.181178e-17 4.712852e-06 3.212503e-06 1.526217e-67
## [11] 3.518016e-08 1.963601e-01 4.894750e-12 5.526564e-43 7.174472e-02
##
## $comparisons
## [1] "Anger - Anxiety"      "Anger - Boredom"      "Anxiety - Boredom"
## [4] "Anger - Happiness"    "Anxiety - Happiness"  "Boredom - Happiness"
## [7] "Anger - Neutral"      "Anxiety - Neutral"    "Boredom - Neutral"
## [10] "Happiness - Neutral"  "Anger - Sadness"      "Anxiety - Sadness"
## [13] "Boredom - Sadness"    "Happiness - Sadness"  "Neutral - Sadness"

```

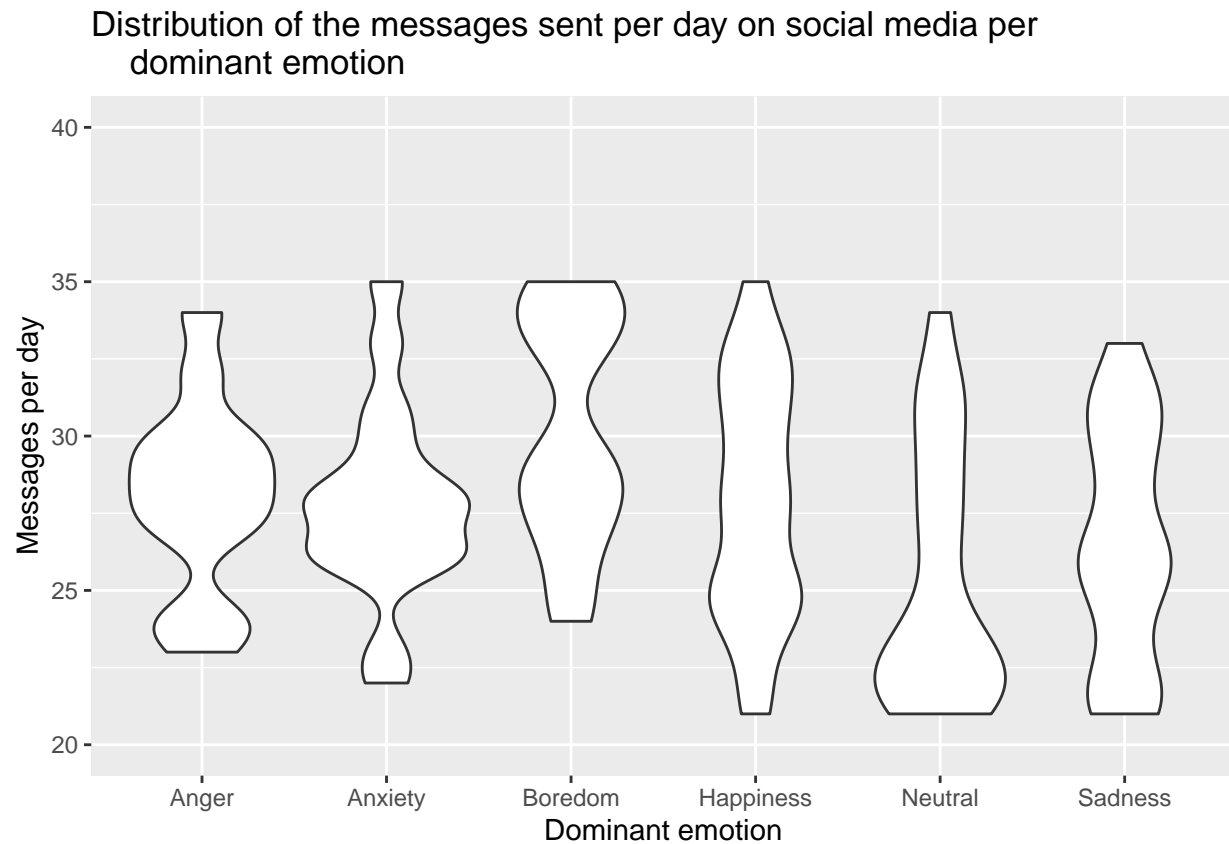
Significant difference = Between both emotions, one of them has more/less age

Considering significance = 0.05 -> CL = 95%:

- "Anger - Anxiety"
- "Anger - Boredom"
- "Anxiety - Boredom"
- "Anger - Happiness"
- "Anxiety - Happiness"
- "Boredom - Happiness"

- “Anger - Neutral”
- “Anxiety - Neutral”
- “Boredom - Neutral”
- “Happiness - Neutral”
- “Anger - Sadness”

```
ggplot(data_transformed, aes(x = Dominant_Emotion, y = Age)) +
  geom_violin() +
  scale_y_continuous(limits = c(20, 40),
    breaks = seq(20, 40, by = 5)) +
  labs(
    title = "Distribution of the messages sent per day on social media per
    dominant emotion",
    x = "Dominant emotion",
    y = "Messages per day")
```



Taking a look at the graphic above, H1 is confirmed.