Assignment 3

Desiree Michel Perez & Toni McCoy

2025-10-27

```
# Assignment 3 - Git Hub Project

# Visualization 2: Creates stratified bar charts of text messages Group and
# Time (Hint: Faceted Bar Charts).

#DM edit: avoid setwd() so the script works on any machine/RCloud
#setwd("~/BHDS 2010/Githubproject")  #removed

#DM edit: foolproof CSV path (works if file is in repo OR in my data/)



csv_path <- if (file.exists("TextMessages.csv")) "TextMessages.csv" else "data/TextMessages.csv"




#DM edit: load required packages up front in the script to ensure it loads
#everytime (self-contained script)
# install.packages("reshape2")  # run once if needed
# install.packages("ggplot2")   # run once if needed
library(reshape2)
library(ggplot2)

# load in the data
txtmssg <- read.csv(csv_path, header = TRUE)

# Check if Group is a factor
is.factor(txtmssg$Group)
```

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

```
txtmssg$Group <- as.factor(txtmssg$Group)
is.factor(txtmssg$Group)
```

```
## [1] TRUE
```

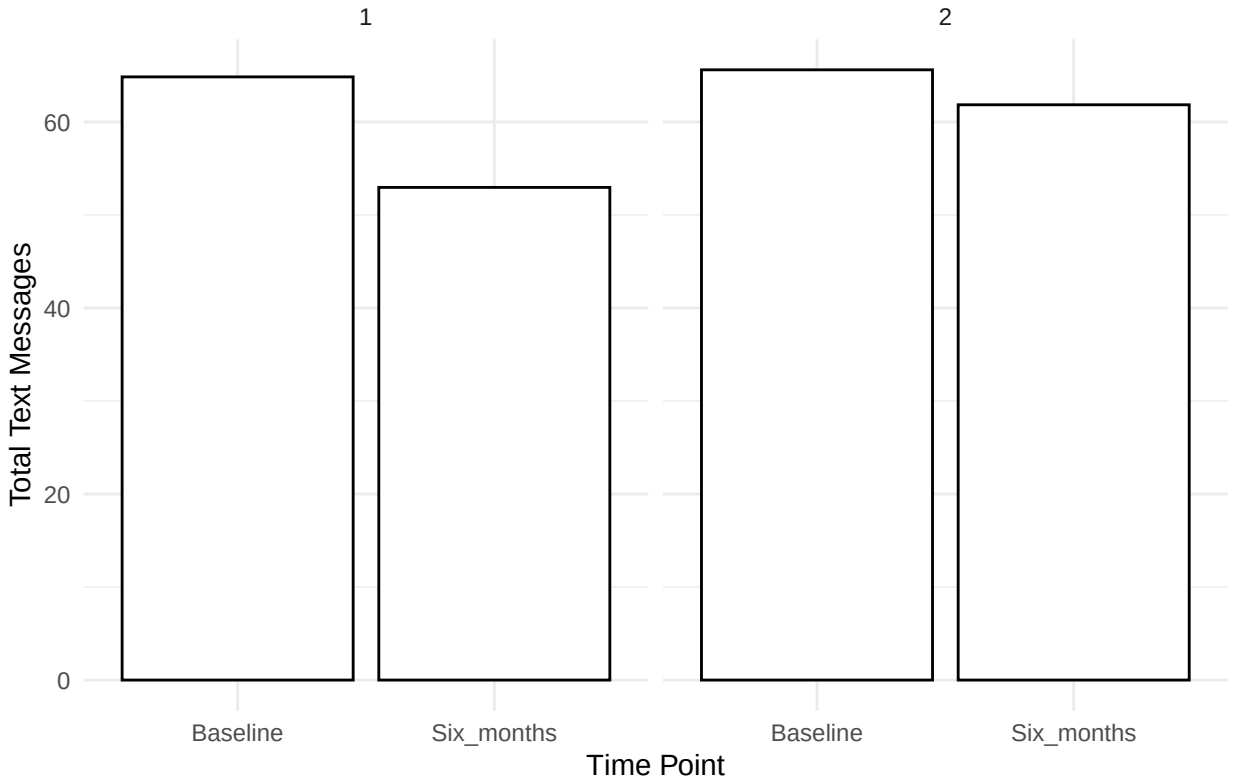
```
# Group is now a factor

# Convert Data to long format
txtmssg_long <- melt(
  txtmssg,
  id.vars       = c("Participant", "Group"),
  variable.name = "Time",
```

```
  value.name    = "Messages"
)

# install packages and load in library for ggplot
# library(ggplot2)  # already loaded above

#DM edit: make plot object explicit so we can print/save all the time
barWithErrors_txtmssg <- ggplot(txtmssg_long, aes(x = Time, y = Messages))

# Check group means to help set plot limits (optional)
by(txtmssg_long$Messages, txtmssg_long$Time, mean)
```

```
## txtmssg_long$Time: Baseline
## [1] 65.22
## ------------------------------------------------------------
## txtmssg_long$Time: Six_months
## [1] 57.4
```

```
# Create a faceted bar plot showing total text messages by time point,
# with red mean points and 95% confidence intervals, faceted by Group
p_bar <- barWithErrors_txtmssg +
  stat_summary(fun = mean, geom = "bar", fill = "white", colour = "black") +
  stat_summary(fun.data = ggplot2::mean_cl_normal, geom = "pointrange", colour = "red") +
  labs(
    title = "Mean Total Text Messages by Time Point For Group 1 and Group 2",
    x = "Time Point",
    y = "Total Text Messages"
  ) +
  facet_wrap(~ Group) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))  # center title

#DM edit: print so plot appears in Plots pane in RCloud
print(p_bar)
```

```
## Warning: Computation failed in `stat_summary()`.
## Computation failed in `stat_summary()`.
## Caused by error in `fun.data()`:
## ! The package "Hmisc" is required.
```




```
1#DM edit: save to a standard figures/ folder (create if missing)
```

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

```
dir.create("figures", showWarnings = FALSE)
ggsave("figures/faceted_bar_means.png", p_bar, width = 8, height = 4.5, dpi = 300)
```

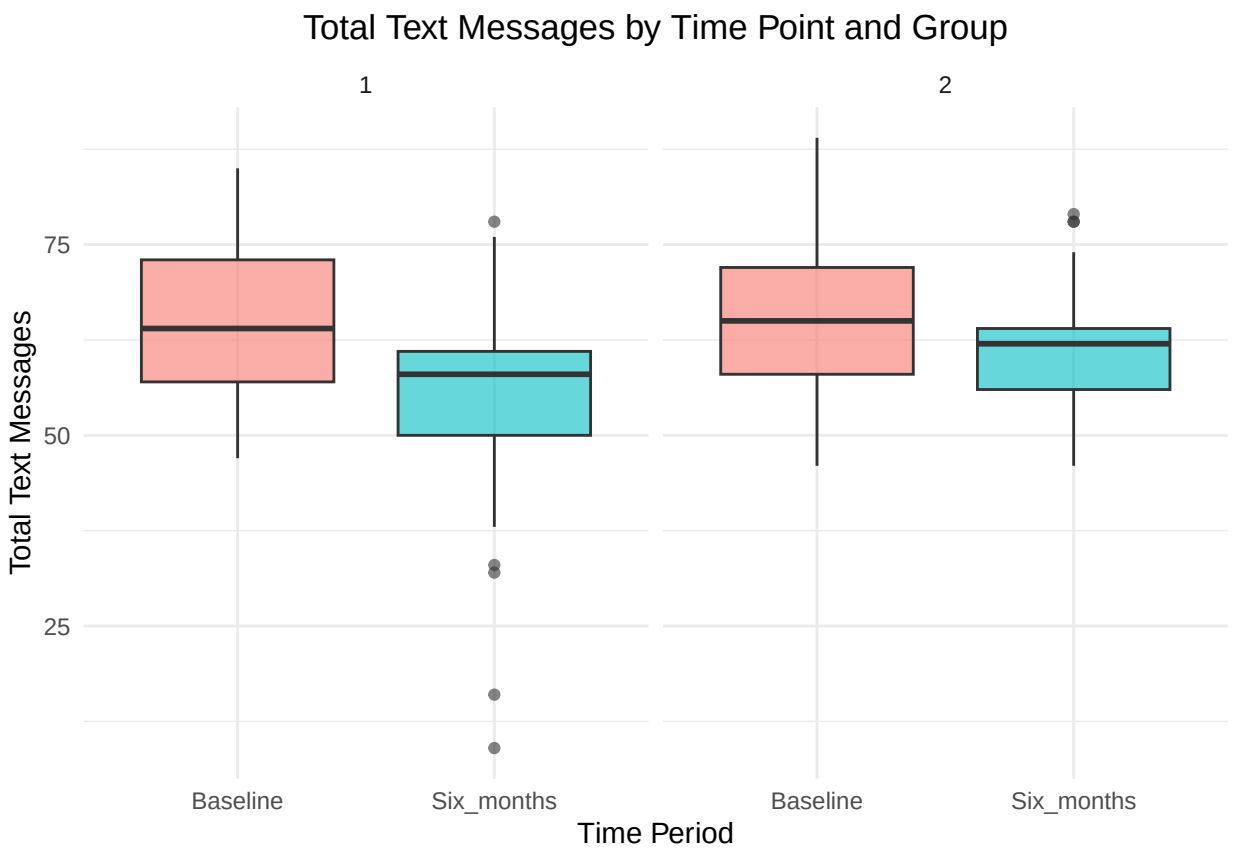
```
## Warning: Computation failed in `stat_summary()`.
## Computation failed in `stat_summary()`.
## Caused by error in `fun.data()`:
## ! The package "Hmisc" is required.
```

```
# The plot shows mean total text messages at two time points (Baseline and
# Six months) for two groups (Group 1 on the left, Group 2 on the right). The
# bars represent the mean, and the red points with error bars likely indicate the
# mean +/- confidence interval (or standard error). In Group 1 (left facet) for
# Baseline the mean total texts are around 65, at Six months the mean total texts
# drop to around 50-55, showing a decrease over time. The error bars overlap
# slightly between Baseline and Six months, suggesting there may be some
# variability, but there's a noticeable downward trend. In Group 2 (right facet)
# The baseline mean total texts are around 65, similar to Group 1, at Six months
# the mean total texts slightly decrease to around 60-62, which is a smaller drop
# than Group 1. Error bars for Group 2 at both time points mostly overlap,
# indicating less pronounced change and possibly more consistency within the group.
# Both groups start at a similar baseline. Group 1 shows a more noticeable
# decrease in text messages over six months, while Group 2 remains relatively
# stable.
```

```r
#DM edit: load required packages up front (self-contained script)
# install.packages("reshape2")  # run once if needed
# install.packages("ggplot2")   # run once if needed
library(reshape2)
library(ggplot2)

# load in the data
txtmssg <- read.csv(csv_path, header = TRUE)

# Check if Group is a factor
is.factor(txtmssg$Group)
```

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

```r
txtmssg$Group <- as.factor(txtmssg$Group)
# Group is now a factor

# Convert Data to long format
txtmssg_long <- melt(
  txtmssg,
  id.vars       = c("Participant", "Group"),
  variable.name = "Time",
  value.name    = "Messages"
)

# Create faceted boxplot object
boxplot_txtmssg <- ggplot(txtmssg_long, aes(x = Time, y = Messages, fill = Time))

# Plot faceted boxplot by Group
p_box <- boxplot_txtmssg +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Total Text Messages by Time Point and Group",
    x = "Time Period",
    y = "Total Text Messages"
  ) +
  facet_wrap(~ Group) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none")

#DM edit: print so the plot appears in the Plots pane
print(p_box)
```




```
#DM edit: save to a standard figures/ folder
dir.create("figures", showWarnings = FALSE)
ggsave("figures/faceted_boxplot_by_group_time.png", p_box, width = 8, height = 4.5, dpi = 300)


# Group 1 has a median text messages decrease from Baseline to Six months, the
# spread (IQR) seems slightly narrower at Six months. Group 1 also has several
# outliers below the lower whisker at six months, indicating a few participants
# sent far fewer messages than the rest. Group 2 has a median text messages
# remain roughly stable from Baseline to Six months. The spread is similar at
# both points in time. There is also only a couple of mild outliers at Six months.
# At Baseline, both groups have similar medians. At Six months, Group 1 shows a
# decrease in median messages, while Group 2 stays consistent.
# Group 1 shows more variability at Six months due to several low outliers, group
# 2 maintains relatively consistent distribution across time.
# Group 1 may have experienced a reduction in text activity over six months,
# while Group 2's behavior remained stable.

library(tidyverse)  #for data wrangling and plotting
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.1.0     v tidyr     1.3.1
## -- 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
```

```
library(broom)      #to clean model outputs
library(leaps)      #for regsubsets
library(tidyr)      #use to reshape data from wide to long
library(dplyr)      #summary stats by group
library(readr)      #change and make cvs

#Im going to also include explicit paths for our three tasks of analysis,
#figures and data wrangling just to make our repository neater
DATA_PATH <- "data/TextMessages.csv"
OUT_ANALYSIS <- "analysis"
OUT_FIGURE <- "figs"
dir.create(OUT_ANALYSIS, showWarnings = FALSE)
dir.create(OUT_FIGURE, showWarnings = FALSE)

#we should always load the data and inspect its structure before any type of
#exploratory analysis or wrangling
text_data <- read.csv("data/TextMessages.csv")
head(text_data)
```

```
##   Group Baseline Six_months Participant
## 1     1       52         32           1
## 2     1       68         48           2
## 3     1       85         62           3
## 4     1       47         16           4
## 5     1       73         63           5
## 6     1       57         53           6
```

```
glimpse(text_data)
```

```
## Rows: 50
## Columns: 4
## $ Group       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Baseline    <int> 52, 68, 85, 47, 73, 57, 63, 50, 66, 60, 51, 72, 77, 57, 79~
## $ Six_months  <int> 32, 48, 62, 16, 63, 53, 59, 58, 59, 57, 60, 56, 61, 52, 9,~
## $ Participant <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
```

```
summary(text_data)
```

```
##      Group        Baseline       Six_months    Participant
##  Min.   :1.0   Min.   :46.00   Min.   : 9.0   Min.   : 1.00
##  1st Qu.:1.0   1st Qu.:57.00   1st Qu.:53.0   1st Qu.:13.25
##  Median :1.5   Median :64.50   Median :60.5   Median :25.50
##  Mean   :1.5   Mean   :65.22   Mean   :57.4   Mean   :25.50
##  3rd Qu.:2.0   3rd Qu.:72.75   3rd Qu.:63.0   3rd Qu.:37.75
##  Max.   :2.0   Max.   :89.00   Max.   :79.0   Max.   :50.00
```

```
#one important thing to notice is that our data is in the wide format- in order
#to make analysis and visualization easier we will change this

library(tidyr)

text_long <- text_data |>
  pivot_longer(
    cols = c(Baseline, Six_months),
```

```
    names_to = "Time",
    values_to = "Text_Messages"
  )
#this is just to check
head(text_long)
```

```
## # A tibble: 6 x 4
##   Group Participant Time       Text_Messages
##   <int>       <int> <chr>              <int>
## 1     1           1 Baseline              52
## 2     1           1 Six_months            32
## 3     1           2 Baseline              68
## 4     1           2 Six_months            48
## 5     1           3 Baseline              85
## 6     1           3 Six_months            62
```

```
write.csv(text_long, "data/TextMessages_long.csv", row.names = FALSE)


#Now that we have changed the dataset from wide to long we will turn our
#attention to the summary statistics of each group and time point with the hope
#of further exploring which variables best explain trends seen. The code below
#is meant to reproduce a table that gives us 4 metrics: N, mean, SD, SE and
#Median. These metrics are imperative to describe data pattern both
#statistically and visually.These metrics were recorded and placed in a csv
#file.

library(dplyr)
library(readr)

summary_tbl <- text_long |>
  group_by(Group, Time) |>
  summarise(
    n = n(),
    mean = mean(Text_Messages, na.rm = TRUE),
    sd = sd(Text_Messages, na.rm = TRUE),
    se = sd / sqrt(n),
    median = median(Text_Messages, na.rm = TRUE),
    .groups = "drop"
  )
print(summary_tbl)
```

```
## # A tibble: 4 x 7
##   Group Time           n  mean    sd    se median
##   <int> <chr>      <int> <dbl> <dbl> <dbl>  <int>
## 1     1 Baseline      25  64.8 10.7   2.14     64
## 2     1 Six_months    25  53.0 16.3   3.27     58
## 3     2 Baseline      25  65.6 10.8   2.17     65
## 4     2 Six_months    25  61.8  9.41  1.88     62
```

```
write_csv(summary_tbl, "analysis/summary_by_group_time.csv")
```