Assignment 3

Anna Hipp Kaplan, Jona Gavazi

2025-11-02

```
################################################################################
# Assignment 3: Collaborating in GitHub.
# Objective: Create shared repository and clone said repository.
# Objective: Assign roles and create branches tailored to Visualization 1
# (Stratified, Faceted Boxplot of Text Messages by Group and Time),
# Visualization 2 (Stratified, Faceted Bar Chart of Text Messages by Group and
# Time), and Summary Statistics (of Text Messages by Group and Time).
# Objective: Update a README.md file with activity and instructions.
################################################################################

# The working directory at hand is the BHDS.2010.Assignments.3 folder, within
# GitHub.

# First, we will want to install the appropriate packages "tidyr" and "ggplot2"
# and then utilize the library() function to load in the packages and their
# dependencies. We also will name an object "text" and read the .csv file
# TextMessages unto it, upon which we will generate our visual summaries
# and run the relevant descriptive statistics.

text <- read.csv("TextMessages.csv")
#install.packages("ggplot2")
library(ggplot2)
#install.packages("tidyr")
library(tidyr)
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
```

Visualization 1

```
# To get a quick understanding of the exact names of the variables involved,
# we can simply run the names() function, followed by the glimpse() command, to
# ensure that the data are in order and accurately reflect the csv file.
names(text)
```

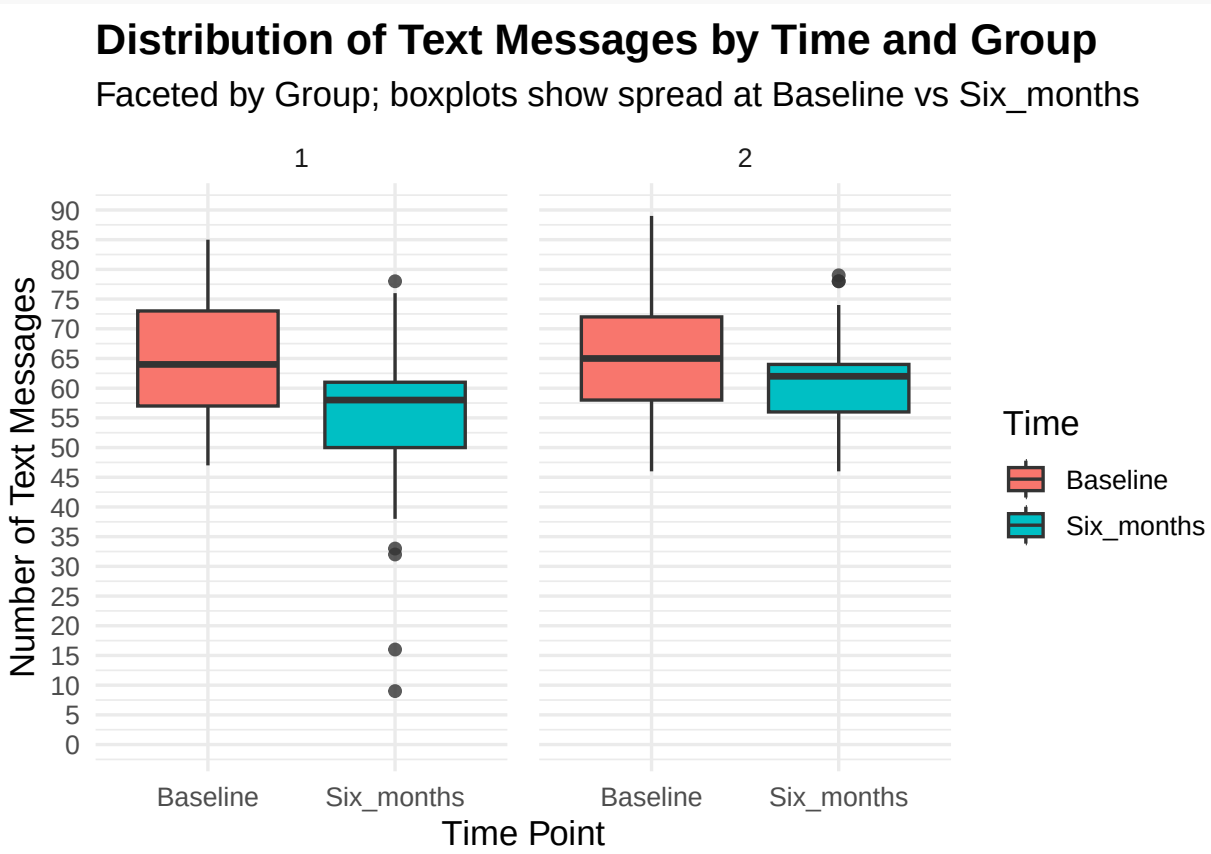
```
## [1] "Group"       "Baseline"    "Six_months"  "Participant"
```

```
glimpse(text)
```

```
## 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,~
```

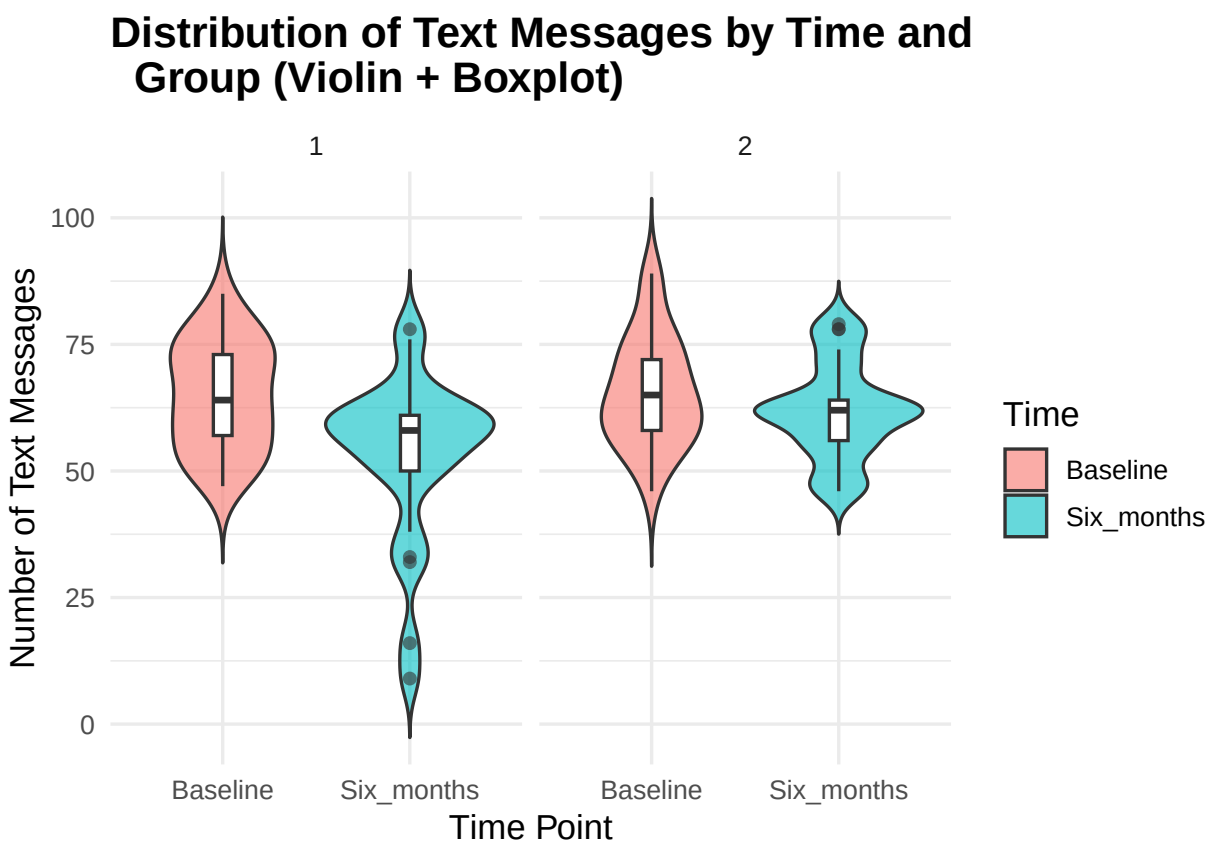
```
# We want to first covert the variables Baseline and Six_months into numeric
# format, to ensure that they can be utilized for quantitative analyses and
# visualizations. We then need to reshape the dataset from wide format to
# long format, using the pivot_longer() command. In the original wide structure,
# each participant featured separate columns for the number of text messages
# sent at Baseline and at Six months. After reshaping, these two columns are
# combined into a single variable called TextMessages, with a corresponding
# Time variable, which indicates whether each observation came from the
# Baseline or Six-month time point. This transformation will serve to make
# the data tidy, which is ideal for plotting/faceting and conducting
# group/time comparisons within ggplot2.

text_long <- text %>% mutate(Baseline = as.numeric(Baseline),
    Six_months = as.numeric(Six_months)) %>% pivot_longer(
    cols = c(Baseline, Six_months),
    names_to = "Time",
    values_to = "TextMessages")

# We want to refine the structure by explicitly converting key variables into
# factors with defined levels. The Time variable (split between indication of
# Baseline or Six months) will be converted into a factor and ordered, so that
# Baseline appears first, ensuring consistency in the plots. The Group variable
# will also be converted into a factor, so that it represents categorical data
# rather than numeric values. In defining these variables as factors, we
# ensure that ggplot2 treats them as categorical axes rather than continuous
# scales.

text_long <- text_long %>% mutate(Time  = factor(Time,
    levels = c("Baseline", "Six_months")),
    Group = as.factor(Group))

# We want to generate a set of faceted boxplots that display the distribution
# of texts across the two time points for each participant group. The function
# ggplot() will map Time to the x-axis, TextMessages to the y-axis, and Time
# to the fill color, to create distinction between the periods. We will then
# utilize the geom_boxplot() layer to create boxplots that show the central
# tendency, variability, and potential outliers. The facet_wrap(~ Group)
# function produces separate panels for each group, allowing for side-by-side
# comparisons for texting behaviors.

boxplot1 <- ggplot(text_long, aes(x = Time, y = TextMessages, fill = Time)) +
  geom_boxplot(outlier.alpha = 0.8) + facet_wrap(~ Group) + labs(
  title = "Distribution of Text Messages by Time and Group",
  subtitle = "Faceted by Group; boxplots show spread at Baseline vs Six_months",
```

```
    x = "Time Point",
    y = "Number of Text Messages",
    fill = "Time") + scale_y_continuous(limits = c(0, 90),
    breaks = seq(0, 90, by = 5)) + theme_minimal(base_size = 13) + theme(
    plot.title = element_text(face = "bold"),
    panel.spacing = unit(12, "pt"))
print(boxplot1)
```




```
# The faceted boxplots illustrate the distribution of text counts, for each
# participant Group (1, 2) across two time points (Baseline, Six_months).
# In both groups, the median number of text messages is moderately high, with
# slightly greater variability at Baseline, as compared to Six months (indicated
# by the wider interquartile range and the presence of a few lower outliers).
# Across groups, the pattern suggests a small decline in text frequency over
# time, with the central tendencies remaining fairly consistent.
# The consistency of medians across groups and the overlapping interquartile
# rages thus indicate that no dramatic behavioral shift occurred, with a rather
# gradual convergence toward more uniform texting patterns over time (as
# indicated by the compactness of the boxes at the Six_months mark). The
# whiskers demonstrate that most of the values fall within a similar overall
# range.

# *AS A BONUS* We can also consider incorporating a violin plot, which expands
# upon the information provided by the faceted box-plot, combining its elements
# with a kernel density plot. This ultimately allows for a more comprehensive
# view of the data's distribution. Box-plots do well to summarize the central
# tendency and spread through medians/quartiles/outliers, but violin plots
```

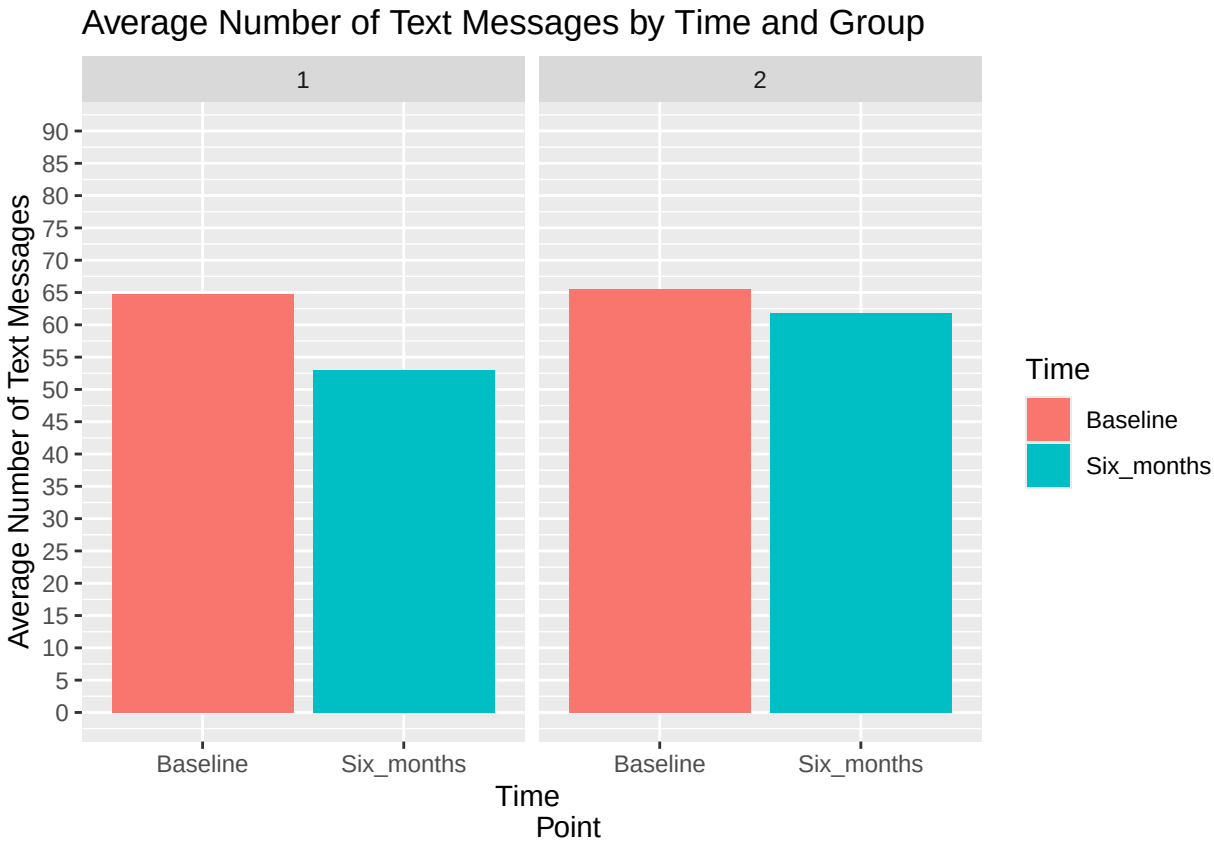
```
# additionally display the shape and density of the data across the
# range of values. This can be especially useful for detecting skewness or
# subtle distributional differences between groups or time points. Within
# ggplot() function, we layer in geom_violin(), still faceted to separate the
# panels by Group.

violinplot <- ggplot(text_long, aes(x = Time, y = TextMessages, fill = Time)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, fill = "white", outlier.alpha = 0.6) +
  facet_wrap(~ Group) + labs(title = "Distribution of Text Messages by Time and
  Group (Violin + Boxplot)", x = "Time Point", y = "Number of Text Messages",
  fill = "Time") + theme_minimal(base_size = 13) + theme(plot.title =
  element_text(face = "bold"))
violinplot
```




```
# The violin plot visualizes the full distribution of text message counts for
# each group. The width of each violin depicts the relative frequency
# of participants within a given text-count (wider areas indicate where more
# participant values are concentrated). In both groups, the violins are
# broader around the mid-range at Baseline, and slightly narrower at Six_months,
# indicating that participants' messaging behavior became more clustered
# around the mean. The thinner tails and smoother contours at Six_months suggest
# reduced variability and fewer extreme variables, particularly in Group 2.
# Overall, the violin plot provides further reinforcement of the subtle decline
# in average texting activity alongside a "tightening" of behavioral
# variability over time.
```

Visualization 2

```
# create stratified bar charts of text messages Group and Time using our
# converted data file text_long from above. We want to generate a set of faceted
# bar charts that display the distribution of texts across the two time points
# for each participant group. Using the function ggplot to map the variable
# Time that we created to the x-axis and the variable TextMessages to the
# y-axis, we generate these bar charts. We also utilized Time to differentiate
# the fill color of each time period (Baseline and Six_months). geom_bar()
# creates bar charts where the height of the bar proportional to the number of
# cases in each group, showing us the average number of text messages sent by
# each group at each time period. Finally, using facet_wrap(~ Group) function
# produces separate panels for each group, allowing us to generate side-by-side
# comparisons.

barchart <- ggplot(text_long, aes(x = Time, y = TextMessages, fill = Time)) +
    geom_bar(stat = "summary", fun = "mean", position = "dodge") +
    facet_wrap(~ Group) +
    labs(title = "Average Number of Text Messages by Time and Group", x = "Time
         Point", y = "Average Number of Text Messages", fill = "Time") +
    scale_y_continuous(limits = c(0,90), breaks = seq(from = 0, to = 90, by =
                                                      5))

# view bar chart
barchart
```



```
# This plot shows the average number of text messages sent in each group, at
# both the baseline and six month mark of the observational period. On the left
# side, we see Group 1, who sent an approximate average of 65 texts at the
# beginning of the observational period and an approximate average of 53 text
# messages sent at the six month mark. On the right side, we see Group 2, who
# also sent an approximate average of 65 text messages at the beginning of the
# observational period, and an approximate average of 62 text messages at the
# six month mark. From this plot, it seems that both groups had a decline in
# the amount of texting they did, with Group 1 having a steeper decline between
# the beginning of the observational period and the six month mark.
```

Summary Statistics

```
# compute summary statistics by Group and Time to examine the number of text
# messages sent by participants in each group at baseline and at six months.
summary_stats <- text_long %>%
  group_by(Group, Time) %>%
  summarise(
    n = n(),
    mean = mean(TextMessages),
    sd = sd(TextMessages),
    min = min(TextMessages),
    max = max(TextMessages))
```

```
## `summarise()` has grouped output by 'Group'. You can override using the
## `.groups` argument.
```

```
# view results
print(summary_stats)
```

```
## # A tibble: 4 x 7
## # Groups:   Group [2]
##   Group Time           n  mean    sd   min   max
##   <fct> <fct>      <int> <dbl> <dbl> <dbl> <dbl>
## 1 1     Baseline      25  64.8 10.7     47    85
## 2 1     Six_months    25  53.0 16.3      9    78
## 3 2     Baseline      25  65.6 10.8     46    89
## 4 2     Six_months    25  61.8  9.41    46    79
```

```
# For Group 1, the mean number of text messages decreased from 64.84
# (SD = 10.68) at baseline to 52.96 (SD = 16.33) at six months, suggesting a
# reduction in texting activity over time. Message counts ranged from 47 to 85
# at baseline and from 9 to 78 at six months, the largest range in our data. For
# Group 2, the mean number of text messages showed a smaller decline, from 65.60
# (SD = 10.84) at baseline to 61.84 (SD = 9.41) at six months. Message counts
# ranged from 46 to 89 at baseline and from 46 to 79 at six months. Overall,
# both groups exhibited a decrease in texting activity over the six-month
# period, with Group 1 seeming to show a larger reduction on average than Group
# 2. This could be a point of further analysis.

################################################################################

# The collective of the findings from the graphics and summary statistics
# present a consistent narrative about participants' texting behaviors over the
```

```
# six month observation period. Both visualizations reveal that text message
# frequency declined slightly from Baseline to Six_months in both groups, with
# Group 1 showing a more pronounced decrease. The summary statistics reinforce
# these patterns, as Group 1's mean text count dropped from ~64.84 to 52.96,
# while Group 2's mean declined more modestly from 65.60 to 61.84. The boxplots
# further illustrate that the variability in texting behavior decreased at the
# Six_month mark, pointing to activity levels becoming more consistent over time.
# The results suggest that while both groups maintained broadly similar
# communication habits, there was a subtle overall reduction in texting activity
# and a convergence toward greater behavioral consistency across the two time
# points, with only mild group-level differences in degree of change.
```

Paired t-test (bonus)

```
# as another bonus, we wanted to conduct a paired t-test to see if the decreasing in texting behavior a

# split data by group
group1 <- text %>% filter(Group == 1)
group2 <- text %>% filter(Group == 2)

# paired t-test for Group 1
t_group1 <- t.test(group1$Baseline, group1$Six_months, paired = TRUE)
t_group1
```

```
##
##  Paired t-test
##
## data:  group1$Baseline and group1$Six_months
## t = 3.3845, df = 24, p-value = 0.002449
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   4.635432 19.124568
## sample estimates:
## mean difference
##           11.88
```

```
# paired t-test for Group 2
t_group2 <- t.test(group2$Baseline, group2$Six_months, paired = TRUE)
t_group2
```

```
##
##  Paired t-test
##
## data:  group2$Baseline and group2$Six_months
## t = 2.0086, df = 24, p-value = 0.05596
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.1035534  7.6235534
## sample estimates:
## mean difference
##            3.76
```

```
# For Group 1, with a p-value = 0.0024, we reject the null hypothesis that the
# true mean difference between the baseline and six month marks is 0 at the
```

```
# alpha = 0.05 significance level. Thus, we can conclude that Group 1 did in fact
# have a statistically significant decline in texts during the observational period.
# We are 95% confident the true mean difference lies between 4.64 and 19.12.

# For Group 2, with a p-value = 0.05596, we fail to reject the null hypothesis
# that the true mean difference between the baseline and six month marks is 0 at
# the alpha = 0.05 significance level. Thus, we can conclude Group 2 did not have
# a statistically significant decline in texts during the observational period.
# We are 95% confident the true mean difference lies between -0.1036 and 7.6236.
# Note that 0 is included in this interval, meaning there is some probability that
# no decline in texting behavior occurs.

# Overall, these results suggest that Group 1 showed a significant reduction in
# texting behavior over time, whereas Group 2's change was smaller and nonsignificant.
```