Assignment 3: Collaborating in Github

Brandon Yee   Veronica Leary

2025-03-24

Introduction

This report generates and displays summary statistics of text message counts by **Group** and **Time Point**. We calculate measures of central tendency and variability to understand texting patterns across two groups and over time.

Load Required Libraries

```r
library(readr)          # Reading CSV files
library(gt)             # Generating clean, styled summary tables
library(ggplot2)        # For plotting
library(dplyr)          # For data manipulation
library(tidyr)          # For reshaping data
library(wesanderson)    # For Wes Anderson-inspired color palettes
library(reshape)        # For converting data to long with melt()

#knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 180), tidy = TRUE)
```

Data Loading and Cleaning

```r
# set working directory as folder on desktop
# setwd("C:/Users/brand/OneDrive/Desktop/BHDS2010/ASSIGN3/bhds-assign-3")
# setwd("/Users/vleary71/Desktop/BHDS2010/ASSIGN3/bhds-assign-3")
# successfully set working directory

# Read in the dataset and clean header rows
data <- read.csv("TextMessages.csv")  # Reads the dataset into an R dataframe
```

```r
# Reshape from wide to long format
data_long <- data %>%
  mutate(across(c(Baseline, Six_months), as.numeric),
         Group = as.factor(Group)) %>%
  pivot_longer(cols = c(Baseline, Six_months),
               names_to = "Time",
               values_to = "TextMessages")
```

Summary Statistics Calculation

We calculate:

- **Count** of observations per group/time
- **Mean**, **Median**, and **Standard Deviation** of text message counts

```r
summary_table <- data_long %>%
  group_by(Group, Time) %>%
  summarise(
    Count = n(),
    Mean = round(mean(TextMessages, na.rm = TRUE), 2),
    Median = round(median(TextMessages, na.rm = TRUE), 2),
    SD = round(sd(TextMessages, na.rm = TRUE), 2),
    .groups = "drop"
  )
```

Summary Table Output

```r
summary_table %>%
  gt() %>%
  tab_header(
    title = "Summary Statistics of Text Messages",
    subtitle = "Grouped by Treatment Group and Time Point"
  ) %>%
  cols_label(
    Group = "Group",
    Time = "Time Point",
    Count = "N",
    Mean = "Mean",
    Median = "Median",
    SD = "Standard Deviation"
  ) %>%
  fmt_number(columns = c(Mean, Median, SD), decimals = 2) %>%
  tab_options(
    table.font.size = 12,
```

<div align="center">

Summary Statistics of Text Messages

Grouped by Treatment Group and Time Point

</div>

Group	Time Point	N	Mean	Median	Standard Deviation
1	Baseline	25	64.84	64.00	10.68
1	Six_months	25	52.96	58.00	16.33
2	Baseline	25	65.60	65.00	10.84
2	Six_months	25	61.84	62.00	9.41

```
    heading.title.font.size = 16,
    heading.subtitle.font.size = 14
)
```

Inference

- If the **mean** and **median** differ substantially, this may suggest skewness in message volume.
- Compare between **Groups** to explore differences in texting behavior.
- An increase from **Baseline** to **Six_months** may indicate behavioral changes over time.
- Use standard deviation to understand variability within each subgroup.

```
###Visualization 1:
#Stratified boxplot of text messages by Group and Time
#Hint: Faceted Boxplot

#Read data set in
#Use read.csv since the file is a csv file
text_data <- read.csv("TextMessages.csv")
#File was successfully read in

#Use nrow() to check the number of rows/observations
nrow(text_data)
```

```
## [1] 50
```

```
#There are 50 rows in the dataset

#Use names() to view the variable names
names(text_data)
```

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

```
#There are variables "Group", "Baseline", "Six_months" and "Participant"

#Using cbind to combine the melted text data without the Group variable with a
#a column containing the Group variable replicated a second time.
long_text_data <- cbind(melt(text_data[,-1],
```

```r
                              id.vars = "Participant", #not melting Participant
                              variable_name = "Time", #Variable name for melted
                    value.names = "Texts"), #argument not working? Supposed
                    #to change the variable name to "Texts", but doesn't
                    #seem to work anymore.
                    Group = rep(text_data$Group, 2)) #Using rep() to replicate

#Use is.factor() to check if Group is a factor
is.factor(long_text_data$Group)
```

[1] FALSE

```r
#FALSE was returned
#Use as.factor() to change it to a factor
long_text_data$Group <- as.factor(long_text_data$Group)
#Verify again with is.factor()
is.factor(long_text_data$Group)
```

[1] TRUE

```r
#TRUE is returned this time

#Check if Time is a factor with is.factor()
is.factor(long_text_data$Time)
```

[1] TRUE

```r
#TRUE was returned

#Check the factor names of Time using levels()
levels(long_text_data$Time)
```

[1] "Baseline"   "Six_months"

```r
#"Baseline" and "Six_months" were returned

#Use levels again and set the names of the factors to have "Six Months" for
#easier readability for the boxplots
levels(long_text_data$Time) <- c("Baseline", "Six Months")
#check the levels again
levels(long_text_data$Time)
```

[1] "Baseline"   "Six Months"

```r
#Now "Baseline" and "Six Months" was returned


######Plot the boxplots

#Use ggplot() with aes set for group on x to stratify and value on y with
```

```
#fill = group to allow the plots to be colored
ggplot(long_text_data, aes(x=Group, y = value, fill = Group)) +
  #adding a boxplot with geom_boxplot()
  geom_boxplot() +
  #Use facet_wrap() to stratify the boxplots by time
  facet_wrap(.~Time) +
  #add labels for the title and y axis
  labs(title = "Boxplots of Text Data Stratified by Group and Time",
       y = "Total Text Messages")+
  #adding a color to the boxplots
  scale_fill_brewer(palette = "Pastel1") +
  #centering the title of the plot
  theme(plot.title = element_text(hjust = 0.5))
```

Boxplots of Text Data Stratified by Group and Time
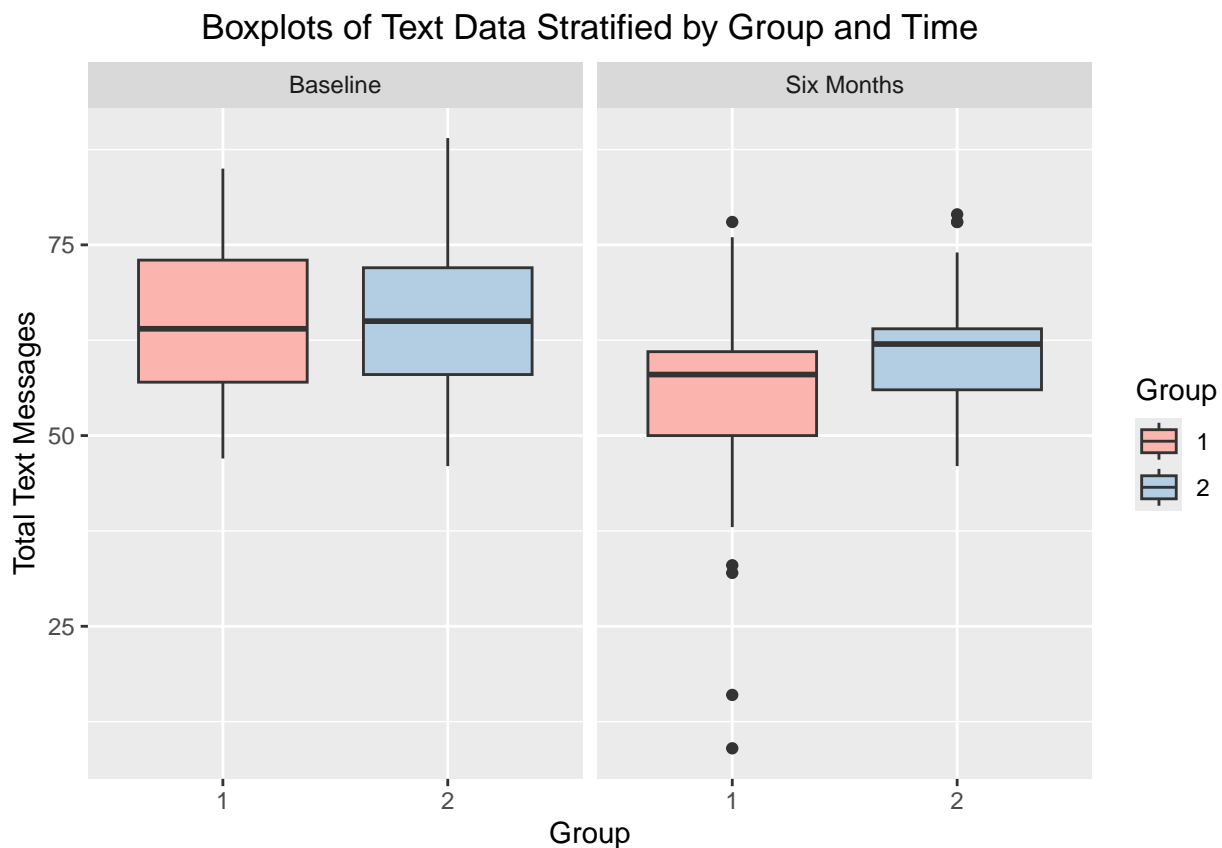

```
#The figure was successfully created
```

```
###Visualization 2:
# stratified_bar_chart.R
# Stratified Bar Chart of Text Messages by Group and Time
# Author: Collaborative GitHub Project Team/Veronica Leary
# Description: This script generates a stratified bar chart with a Wes Anderson color palette using ggp

# Load and clean the dataset
data <- read.csv("TextMessages.csv")  # Load dataset

# Rename columns for clarity
```

```r
colnames(data) <- c("Group", "Baseline", "Six_months", "Participant")

# Remove redundant header row
data <- data[-1, ]

# Convert data types and reshape
data <- data %>%
  mutate(across(c(Baseline, Six_months), as.numeric),
         Group = as.factor(Group)) %>%
  pivot_longer(cols = c(Baseline, Six_months),
               names_to = "Time",
               values_to = "TextMessages")
```

```r
#Use ggplot() with pipeline and dplyr set as groups 1 and 2 and stratified by two points in time-at bas
data %>%
  group_by(Group, Time) %>%
  summarise(TotalMessages = sum(TextMessages, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = Time, y = TotalMessages, fill = Group)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ Group) +
  scale_fill_manual(values = wes_palette("Rushmore", n = 2)) +
  labs(title = "Stratified Bar Chart of Text Messages by Group and Time",
       x = "Time Point",
       y = "Total Text Messages") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```


Stratified Bar Chart of Text Messages by Group and Tir
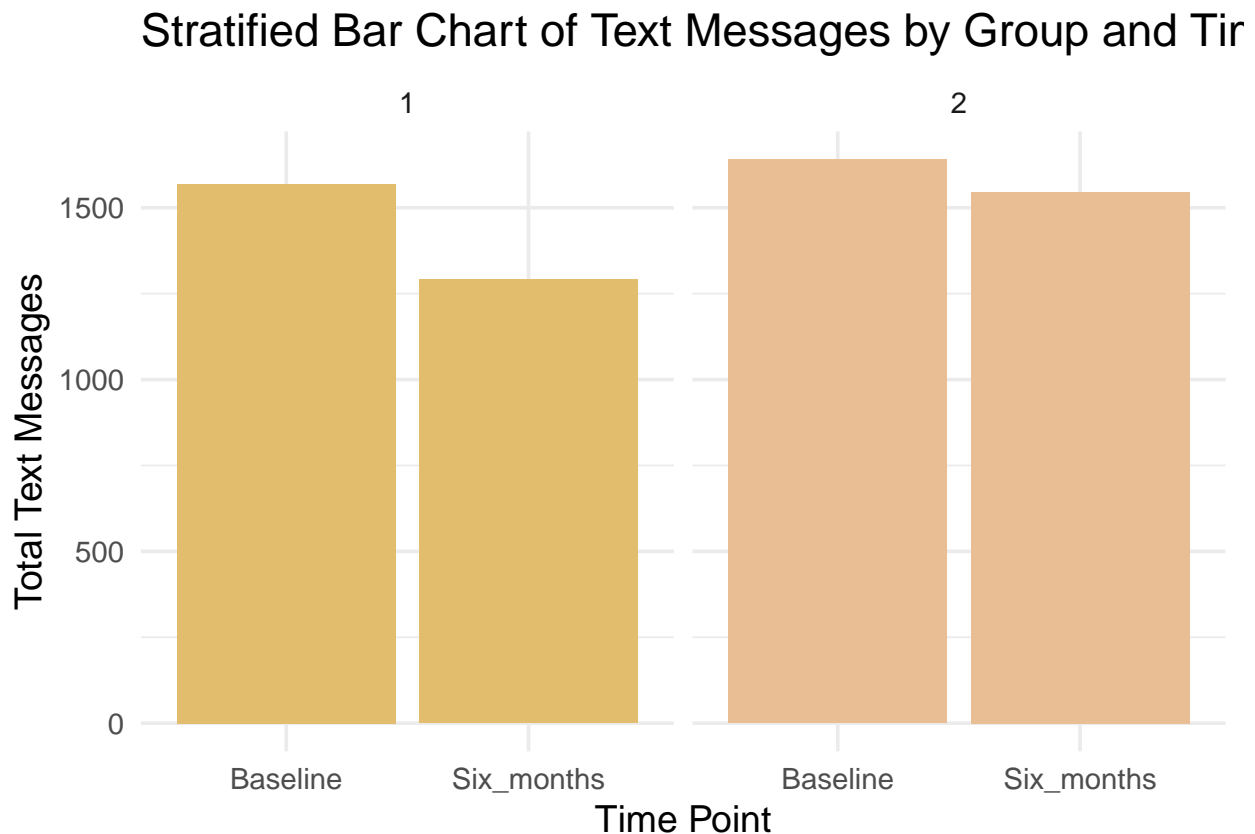
Project Summary

This script outlines the contributions and workflow from our group project analyzing text message data.

It can be used as a reference in combination with the visual and statistical output scripts.

Contributions Overview

Brandon Yee:

- Responsible for Visualization 1: Stratified boxplot using ggplot2 default theme.

- This visualization highlighted the distribution of text messages across time and group,

including medians, variability, and outliers.

Veronica Leary:

- Responsible for Visualization 2: Stratified bar chart using ggplot2 + wesanderson theme.

- Allowed for comparison of total message counts between groups and time points.

- Revealed possible increase in message volume in Group B over time.

- Responsible for Summary Statistics:

- Used dplyr + gt to produce a well-formatted table of N, Mean, Median, and SD.

- Supported the interpretation of patterns observed in the plots.

- Responsible for all documentation:

- Embedded interpretation and narrative into visualizations.

- Created markdown and final report files for submission.

GitHub Workflow

- Created a dedicated branch for visualizations and documentation tasks.

- Commit messages included:

- "Added stratified bar chart with Wes Anderson color palette"

- "Generated summary statistics table with gt"

- "Created documentation and embedded inference blocks"

- Pushes were successful, but merge is pending due to repository permissions

(partner is the owner of the GitHub repo).

Reflection

This assignment helped reinforce:

- The value of clear commit messages and reproducible code.

- Collaborative coding practices using Git and GitHub.

- Communicating visual and statistical insights clearly through embedded narrative.