Algebraic Thinking in Primary School with Systems Thinking Tools

Alime Tunalıer[a], Meltem Ceylan Alibeyoğlu[b1], Emre Göktepe[c]

[a] Alime Tunalıer, Ministry of Education, Bursa, Türkiye
[b] Meltem Ceylan Alibeyoğlu, PhD, Darüşşafaka Schools, İstanbul, Türkiye
[c] Emre Göktepe, Systems Thinking Association, İzmir, Türkiye

Keywords

Systems thinking, problem solving, classroom implementation, systems literacy, algebraic thinking

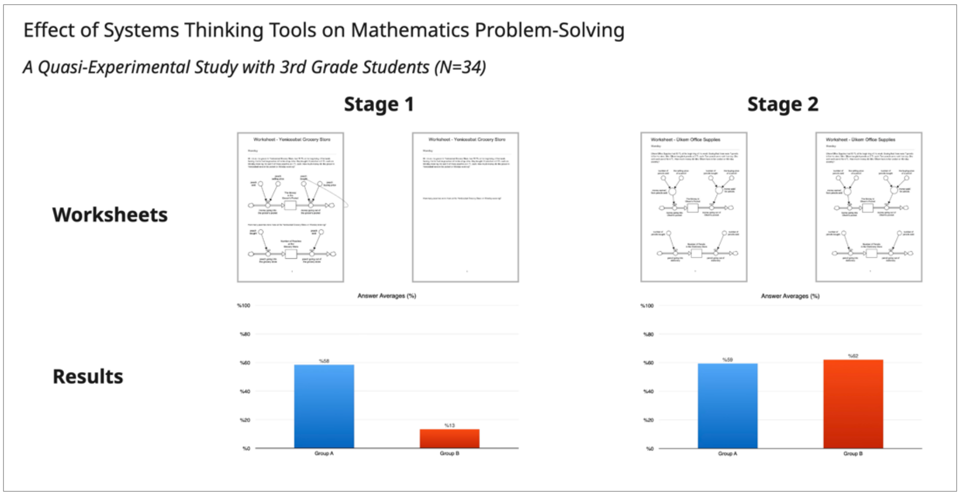
STRUCTURED ABSTRACT

Introduction to the Problem:
Traditional mathematics education often emphasizes procedural skills over conceptual understanding, limiting students' ability to recognize relationships and reason about systems. The Turkish National Curriculum emphasizes holistic, process-oriented learning but lacks concrete strategies for implementing systems thinking in classrooms.
Approach to the Work:
This action research study investigated the effect of a systems thinking tool —stock-flow diagram— on third-grade students' ability to solve scenario-based math problems. A quasi-experimental design compared two groups: one with prior exposure to systems tools and one without.

Results:
In the initial stage, the experienced group outperformed the inexperienced group significantly ($p = 0.000$). After a short introduction to systems tools, the performance gap disappeared ($p = 0.474$). Group B improved significantly ($p = 0.000$) upon adopting the method, while Group A maintained consistent performance ($p = 0.776$).

Discussion:
The findings demonstrate that systems thinking tools can enhance students' problem-solving and algebraic reasoning skills, even with brief exposure. Integrating these tools into mathematics instruction supports curriculum goals and helps students understand relationships and dynamic processes. The study offers a scalable method for fostering systems literacy in early education.

Visual abstract



1 Correspondence author.
E-mail address: meltem.alibeyoglu@dek.k12.tr (Meltem Ceylan Alibeyoğlu).

1 | Introduction

Mathematics education plays a fundamental role in developing students' ability to understand the world, recognize patterns, and think logically. However, conventional instructional methods frequently prioritize procedural fluency over conceptual understanding, resulting in students concentrating on identifying the "correct answer" rather than examining the structure and behavior of problems (Boaler, 2002). This tendency can inhibit students' ability to recognize the interconnectedness of variables in dynamic situations, which is a critical skill for effective real-world problem solving.

One promising response to this issue is the development of quantitative reasoning, which involves making sense of relationships among quantities in a wide range of contexts. It is considered a foundational dimension of mathematical development and provides the conceptual basis for students to engage in algebraic reasoning (Kaput et al., 2008, pp. 3, 98, 126). Research shows that visualization enhances algebraic reasoning after foundational knowledge is established. Prompting students to construct their own visual representations can reduce cognitive load and support more effective problem solving (Ünal et al., 2023).

"Systems thinking is a set of synergistic analytic skills used to improve the capability of identifying and understanding systems, predicting their behaviors, and devising modifications to them in order to produce desired effects. These skills work together as a system" (Arnold & Wade, 2015, p. 8). System dynamics, a formal methodology for systems thinking, utilizes tools such as stock-flow diagrams to model accumulations, rates of change, and feedback processes (Forrester, 2009). These visual representations help learners externalize and explore abstract relationships, making them especially valuable for developing early algebraic reasoning and systems literacy (English & Lehmann, 2024). In recent years, research has demonstrated that even young learners can grasp fundamental systems thinking concepts when supported with developmentally appropriate tools (Feriver et al., 2019, p. 13-15; Koçulu, 2024, p. 233-243; Ünsal, 2024, p. 69-74). The integration of systems thinking tools into early mathematics instruction represents a promising pedagogical innovation. These tools allow students to visualize not only quantities and operations, but also the structure and dynamics of the problem situations, supporting both mathematical modeling and causal reasoning

The Turkish National Curriculum aims to nurture individuals who are not only knowledgeable but also able to connect information, understand processes, and become system-literate individuals (Ministry of National Education [MEB], 2024). In this context, one of the most distinctive aspects of the curriculum is its recommendation to incorporate systems thinking skills in lesson plans. However, the methods, techniques, and tools for implementing systems thinking skills have not yet been defined. At this point, stock-flow diagram, which is a basic system dynamics tool, stand out as an easy-to-understand, easy-to-implement, and scalable option for addressing this gap (English and Lehmann, 2024; Fisher and Systems Thinking Association [STA], 2023). This paper presents an implementation example involving the use of systems thinking tools in a third-grade mathematics class. The effects of the application on students are then evaluated, and the contribution of systems thinking is discussed. Thus, a method for systems thinkin implementation is proposed.

This paper proposes a practical approach to integrating systems thinking tools—specifically stock-flow diagrams—into third-grade mathematics instruction. By demonstrating an application of this accessible, visual tool, this study aims to illustrate how systems thinking can enhance students' quantitative reasoning, conceptual understanding, and algebraic thinking. Furthermore, the paper evaluates the impact of this instructional method on students' ability to visualize, interpret, and reason about dynamic mathematical relationships, contributing significantly to addressing the current pedagogical gap identified in the Turkish National Curriculum (MEB, 2024).

2 | Purpose

The main purpose of this study is to investigate the impact of integrating systems thinking tools—specifically stock-flow diagrams—into third-grade mathematics instruction on students' abilities to solve scenario-based problems. Through this investigation, the paper provides a concrete example of how systems thinking skills can be effectively adapted and implemented within the context of the Turkish National Curriculum (MEB, 2024).

3 | Method

This study employs the action research methodology (Mertler, 2014). Following this approach, the classroom teacher, supported by the Systems Thinking Association, developed an instructional plan to enhance her teaching methods, implemented it, and then observed and evaluated the outcomes.

The research was conducted with third-grade students at a public school located in the Osmangazi district of Bursa province. A total of 34 students participated, divided equally between two classes labeled as Group A and Group B. Group A had experienced inconsistent teaching continuity with four different teachers over previous years, while Group B benefited from consistent instruction by the same teacher for three consecutive years.

The study was structured using a two-stage quasi-experimental design. In the first stage, both groups were presented with the "Yeniceabat Grocery Store" scenario. Group A was provided worksheets integrating stock-flow diagrams while Group B worked on identical questions without these tools. The scenario focused on analyzing peach buying and selling transactions, with students evaluating income-expense balances, stock changes, and price variations. Group A utilized systems thinking tools for calculations and data interpretation, whereas Group B employed conventional calculation methods.

In the second stage, both groups engaged with a similar scenario, "Ülkem Office Supplies," after Group B received a preliminary one-hour session introducing them to systems thinking tools. Both groups then used identical worksheets featuring stock-flow diagrams.

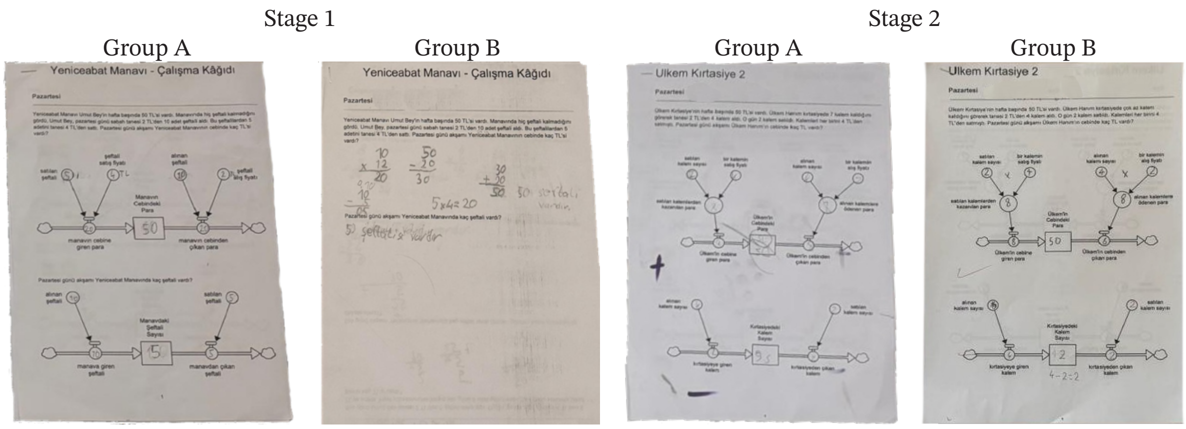


FIGURE 1 | The first pages of the worksheets studied by groups

Statistical analyses were conducted in R (version 4.3.1; R Core Team, 2023). The implementation code was developed with assistance from DeepSeek-R1 (DeepSeek, 2024), an AI language model. The AI system was used to generate initial code structures for non-parametric tests, correlation analyses, and visualizations, which were subsequently verified, modified, and executed by the researchers. The AI contributed to technical implementation but had no role in research design, data interpretation, or conceptual development. Full analysis code is available in Appendix E.

4 | Results

Before comparing the statistics, the Shapiro-Wilk Normality Test was used to investigate whether the data distributions were normally distributed.

TABLE 1 | Shapiro-Wilk Normality Test results

Group	Stage	p_value	Note
1	Stage 1	0.006	Not normally distributed
1	Stage 2	0.201	Normally distributed
2	Stage 1	0.006	Not normally distributed
2	Stage 2	0.070	Normally distributed

As can be seen in Table 1, the Stage 1 data are not normally distributed, while the Stage 2 data are normally distributed. Therefore, different statistical tests were applied:

- Stage 1 - Intergroup Comparison:
  - Purpose: To test whether the initial scores of Group A and Group B are different.
  - Test type: Unpaired test (Group A and Group B consist of different participants).
  - Selected test: Mann-Whitney U test (At least one group is not normally distributed).
- Stage 2 - Intergroup Comparison:
  - Purpose: To test whether the scores of Group A and Group B are different after using the stock-flow diagram.
  - Test type: Unpaired test.
  - Selected test: Independent samples t-test (Both groups are normally distributed).
- Intra-group comparison for Group B:
  - Purpose: To test whether Group B shows a significant change when adopting the stock-flow diagram method.
  - Test type: Paired test (measurements of the same participants at two different times).
  - Selected test: Mann-Whitney U test (At least one measurement is not normally distributed).
- Intra-group comparison for Group A:
  - Purpose: To test whether Group A shows a significant change over time despite using the same method
  - Test type: Paired test
  - Selected test: Mann-Whitney U test (At least one measurement is not normally distributed)

TABLE 2 | Statistical Test Results

Test	Type	Description	Statistics*	p-value	Method	Significance
Stage 1 (Group A & Group B)	Intergroup	Whether the groups were different at the beginning	289	0.000	Mann-Whitney U test	Yes
Stage 2 (Group A & Group B)	Intergroup	Whether the groups are different after using the stock flow diagram	-0.725	0.474	Welch's two-sample t-test	No
Group B (Stage 1 & Stage 2)	Intra-group	Whether Group B improved using the stock flow diagram method	153	0.000	Mann-Whitney U test	Yes
Group A (Stage 1 & Stage 2)	Intra-group	Whether Group A changed over time	47.5	0.776	Mann-Whitney U test	No

*Statistic: t-value for Welch's t-test (measures standardized difference between group means); W-statistic for Mann-Whitney U test (based on sum of ranks). Larger absolute values indicate stronger evidence against the null hypothesis.

Student performance was assessed by counting the number of correct computational answers on the worksheets. Each worksheet contained calculation problems involving addition, subtraction, and multiplication related to grocery store scenarios. A response was scored as correct only if the final numerical result of the calculation was accurate. Because the tasks involved basic arithmetic operations, there were no instances of partially correct answers — calculations were either entirely correct or incorrect.

Results of comparisons (Table 2)

- *Intergroup Comparison (Stage 1):* A substantial difference was identified between Groups A and B in the preliminary stage ($p = 0.000$). Group A, which had prior experience with stock-flow diagrams, demonstrated significantly greater computational accuracy than Group B, which used conventional calculation methods. It was observed that students in Group A demonstrated a higher level of proficiency in solving calculation problems when working with the Yeniceabat Grocery Store worksheet. This outcome indicates Group A's capacity to employ the visual stock-flow diagram structure to methodically track and calculate money flows and inventory changes. In contrast, Group B students employed solely traditional arithmetic methods without the benefit of this structured approach.

- *Intergroup Comparison (Stage 2):* In the second stage, no significant difference was found between the groups (p = 0.474). Following the administration of a one-hour introductory session on stock-flow diagrams to Group B, both groups attained comparable levels of computational accuracy on the Ülkem Office Supplies worksheet. The difference in performance between the two groups that was observed in Stage 1 disappeared, indicating that both groups were equally successful in solving the calculation problems with access to the same stock-flow diagram tools.
- *Intra-Group Change (Group B):* Group B showed significant improvement after adopting the stock-flow diagram method (p = 0.000). A significant increase in the number of correct calculations was observed from Stage 1 to Stage 2. This finding indicates that the method improved Group B's capacity to execute precise arithmetic operations by providing a more structured framework for organizing and monitoring numerical relationships in problem-solving scenarios.
- *Intra-group Change (Group A):* As expected, Group A demonstrated consistent performance over time (p = 0.776) by continuing to apply the stock-flow diagram method throughout both stages. The number of accurate calculations remained consistent across stages, thereby confirming that Group A maintained computational precision by employing the same approach for both problem scenarios.

The substantial disparity observed in Stage 1 was no longer statistically significant in Stage 2, suggesting that the stock-flow diagram method effectively increased Group B's computational precision, aligning it with Group A's performance level. The improvement demonstrated by Group B upon adopting the stock-flow diagram method validates the efficacy of this visual tool in facilitating precise mathematical calculations in scenario-based problems. Group A's consistent performance across both stages provides evidence of the sustained effectiveness of the stock-flow diagram approach.

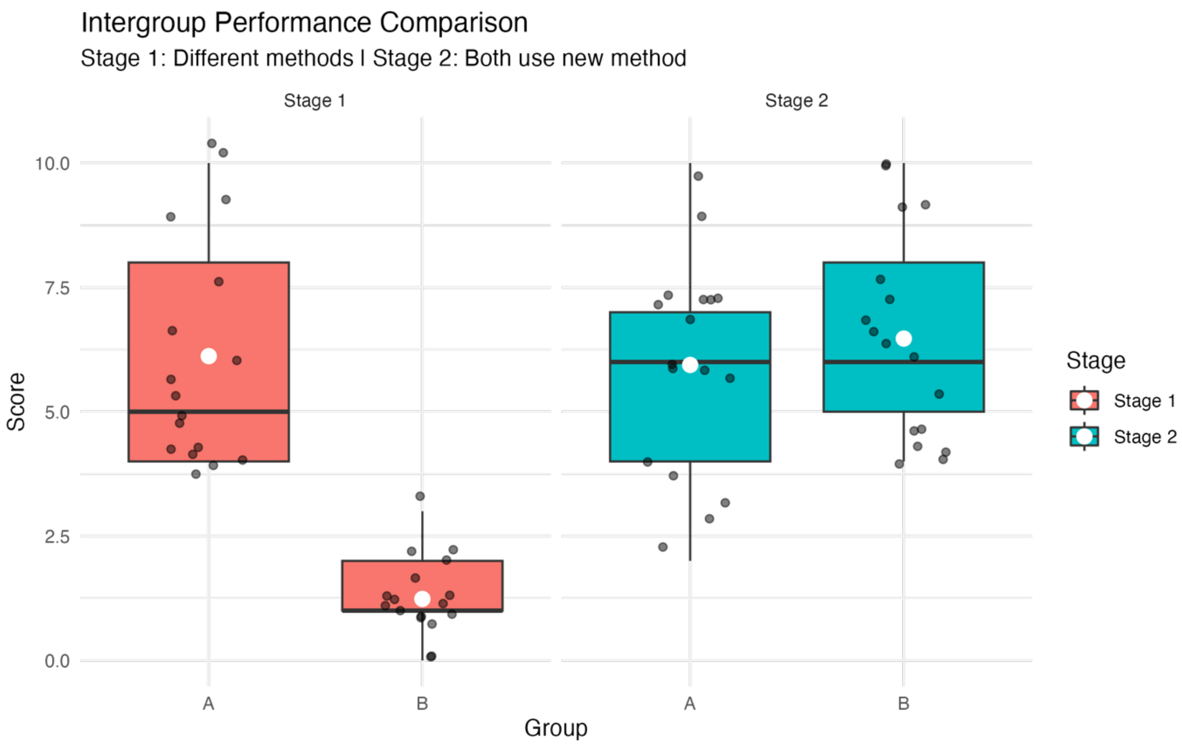


FIGURE 2 | Intergroup performance comparison
(In the box plot, the box [colored rectangle] represents the area containing 50% of the data, the horizontal line inside the box represents the median, the white circle represents the mean, and the vertical lines outside the box extend to the furthest data points that fall within 1.5 × Interquartile Range from the box edges)

This study has revealed that elementary school students are more successful in solving scenario-based problems when they are introduced to systems thinking tools.

The Turkish National Curriculum aims to raise children to be "virtuous, self-actualizing, and productive individuals," emphasizing principles such as a holistic approach, meaning construction, and awareness development. Systems thinking practices are also consistent with these principles. This approach, implemented in mathematics lessons, goes beyond learning based solely on numerical operations, contributing to the development of individuals who establish connections, seek contexts, and understand causality.

5 | Suggestions

This study demonstrates that systems thinking can be acquired at an early age through the use of systems thinking tools (e.g., stock-flow diagrams), in line with the vision of the Turkish National Curriculum. It thus provides evidence that this approach can contribute to both improving the quality of education and developing meaningful skills that individuals can use throughout their lives.

- Systems thinking tools should be actively used not only in science and social studies but also in mathematics education.
- Examples of systems thinking should be increased in in-service training for teachers.
- Skills such as graphic literacy and stock-flow diagram interpretation should be clearly included in teaching programs.
- Learning scenarios should be developed that enable students to discover systemic relationships using examples from their own lives.
- System literacy should be included in teaching programs in a way that incorporates tools such as stock-flow diagrams.

References

Arnold, R. D., & Wade, J. P. (2015). A definition of systems thinking: A systems approach. *Procedia computer science*, *44*, 669-678.

Boaler, J. (2002). Experiencing School Mathematics: Traditional and Reform Approaches to Teaching and Their Impact on Student Learning (2nd ed.). Routledge.

DeepSeek. (2024). *DeepSeek-R1: AI-assisted R code development for statistical analysis (Version 2024-06-19)* [Large language model]. https://www.deepseek.com

English, L. D., & Lehmann, T. (2024). Systems thinking in the early grades. In *Ways of Thinking in STEM-based Problem Solving* (pp. 107-122). Routledge.

Feriver, Ş., Olgan, R., Teksöz, G., & Barth, M. (2019). Systems thinking skills of preschool children in early childhood education contexts from Turkey and Germany. *Sustainability*, *11*(5), 1478.

Fisher, D. M., & Systems Thinking Association. (2023). Systems thinking activities used in K-12 for up to two decades. In *Frontiers in education* (Vol. 8, p. 1059733). Frontiers Media SA.

Forrester, J. W. (2009). Some basic concepts in system dynamics. *Sloan School of Management*, 1-17.

Kaput, J. J. (1999). Teaching and Learning a New Algebra 1. In *Mathematics classrooms that promote understanding* (pp. 133-155). Routledge.

Koçulu, A. (2024). *Development and implementation of the sustainable development goals unit: Exploring students' systems thinking skills and ethical reasoning* [Doctoral dissertation, Yıldız Technical University].

Milli Eğitim Bakanlığı (2024). Türkiye Yüzyılı Maarif Modeli Öğretim Programları Ortak Metni. MEB. Ankara.

R Core Team (2023). _*R: A Language and Environment for Statistical Computing*_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ünal, Z. E., Ala, A. M., Kartal, G., Özel, S., & Geary, D. C. (2023). Visual and Symbolic Representations as Components of Algebraic Reasoning. *Journal of Numerical Cognition*, *9*(2), 327-345.

Ünsal, E. (2024). *The effect of solving environmental problems with systems thinking tools on eighth-grade students' linear equations achievement and system thinking comprehension* [Master's thesis, Bahçeşehir University].

Appendix A
Lesson plan for Yeniceabat Grocery Store (stock-flow diagrams)

3rd Grade - Unit 3 - Numbers and Operations

M.3.1.4. Multiplication with Natural Numbers

Terms and concepts: multiplication, multiplication table, multiplier, product

Symbols: x

M.3.1.4.1. Explains that multiplication is repeated addition.

Includes work with real objects.

M.3.1.4.2. Performs multiplication operations with natural numbers.

a) The meaning of the multiplication symbol (x) is emphasized.

b) Numbers up to 10 are multiplied by 1, 2, 3, 4, and 5.

c) It is emphasized that changing the order of the factors does not change the product.

ç) A multiplication table up to 5 (including 5) is created using a hundred chart and operation tables.

d) The effect of 1 and 0 in multiplication is explained.

M.3.1.4.3. Solves problems that require multiplication with natural numbers.

Problems requiring a single operation are studied.

M.3.1.4.6.Solves problems that require two operations, one of which is multiplication.

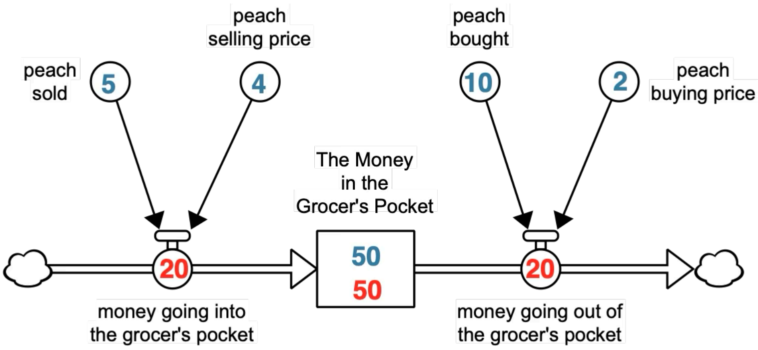
Work on formulating problems is included.

MAT.3.4.1. Ability to work with a single data set based on quantitative data obtained through categorization and counting, and make data-driven decisions.

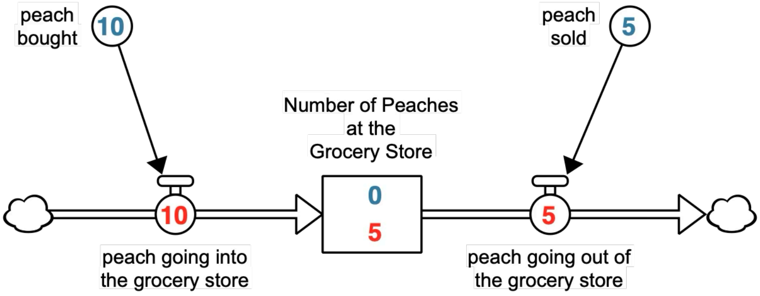
Analyzes data by visualizing it with the tools he chooses.

Monday

Mr. Umut, the grocer in Yeniceabat Grocery Store, had 50 TL at the beginning of the week. Seeing that he had no peaches left in his shop, Umut Bey bought 10 peaches at 2 TL each on Monday morning. He sold 5 of these peaches at 4 TL each. How much money did the grocer in Yeniceabat have in his pocket on Monday evening?

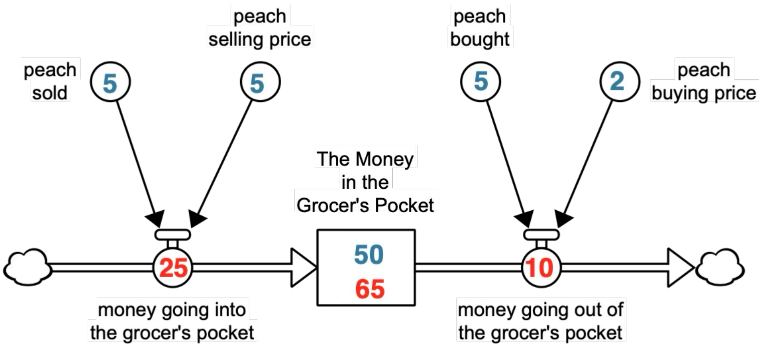


How many peaches were there at the Yeniceabat Grocery Store on Monday evening?

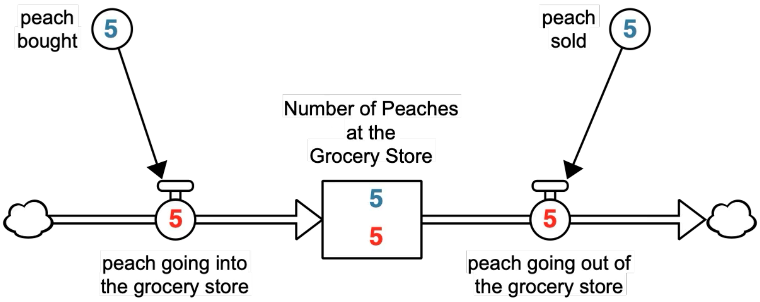


Tuesday

On Tuesday, Mr. Umut bought 5 more peaches at 2 TL each. That day, he raised the selling price of peaches from 4 TL to 5 TL. Despite the price increase, he sold 5 peaches on Tuesday. How much money did Mr. Umut have at the end of Tuesday?

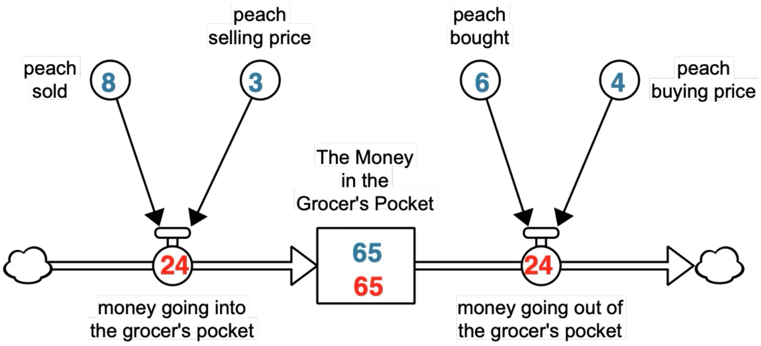


How many peaches were there at the Yeniceabat Grocery Store on Tuesday evening (taking into account the peaches left over from the previous day)?

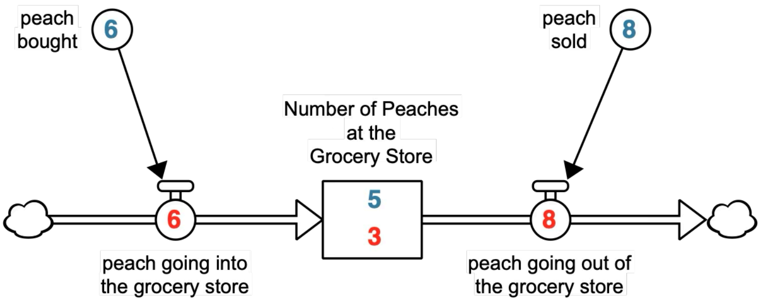


Wednesday

Mr. Umut, owner of Yeniceabat Grocery Store, went to buy fresh peaches on Wednesday and found that the price had increased. The purchase price of peaches was 4 TL. A little disappointed, he bought six peaches at 4 TL each. Despite the increase in the purchase price, he did not raise the selling price; on the contrary, he lowered it from 5 TL to 3 TL. On Wednesday, he sold 8 peaches at 3 TL each. How much money did Mr. Umut have on Wednesday evening?

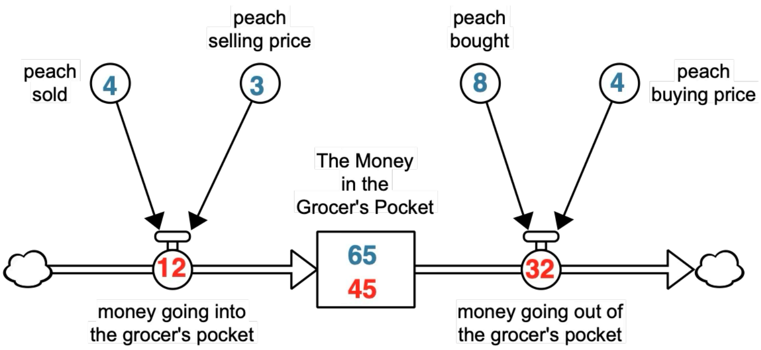


How many peaches were there at the Yeniceabat Grocery Store on Wednesday evening (including peaches left over from the previous day)?

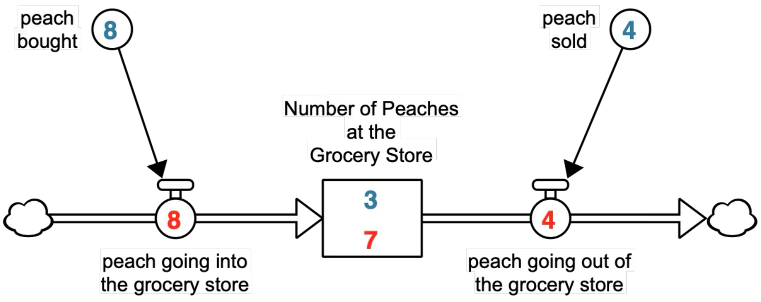


Thursday

On Thursday, thinking that sales would increase, he bought 8 more peaches. He paid 4 TL for each peach, as before. However, sales did not go as he had hoped. He was only able to sell 4 peaches at 3 TL each. How much money did the Yeniceabat Grocery Store have at the end of Thursday?

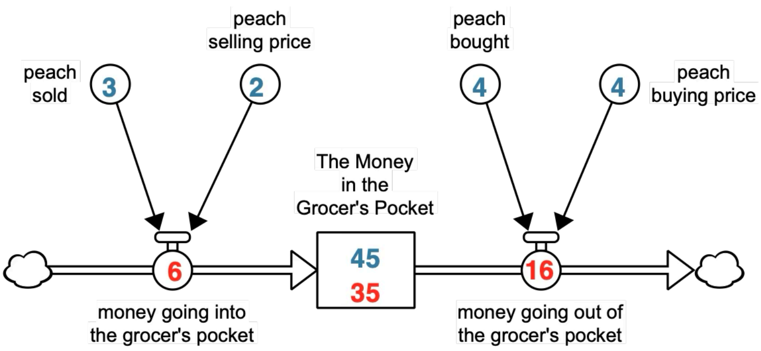


How many peaches were there at the Yeniceabat Grocery Store on Thursday evening (taking into account the peaches left over from the previous day)?

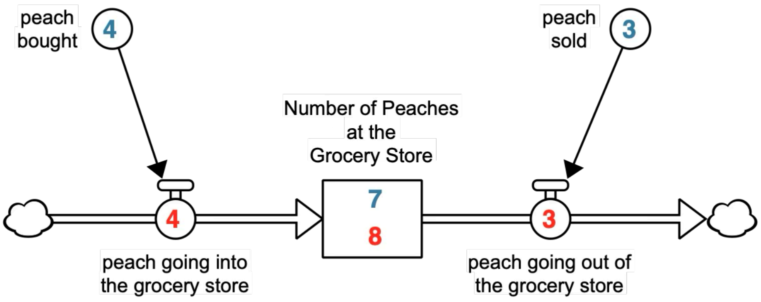


Friday

On Friday morning, he bought four more peaches at 4 TL each. He had decided to lower the selling price even further. On Friday, he would sell the peaches at 2 TL each. But sales fell again. On Friday, he was only able to sell three peaches. How much money did Mr. Umut have left in his pocket at the end of Friday?



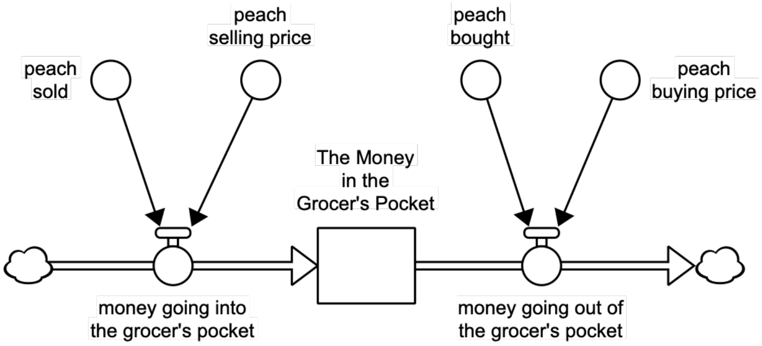
How many peaches were there at the Yeniceabat Grocery Store on Friday evening (taking into account the peaches left over from the previous day)?



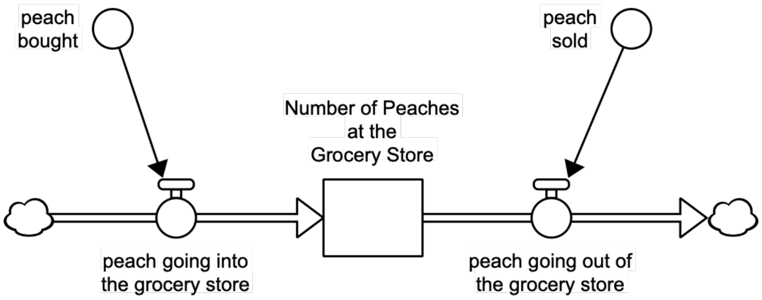
Appendix B
Worksheet for Group A- Yeniceabat Grocery Store (first page)

Monday

Mr. Umut, the grocer in Yeniceabat Grocery Store, had 50 TL at the beginning of the week. Seeing that he had no peaches left in his shop, Umut Bey bought 10 peaches at 2 TL each on Monday morning. He sold 5 of these peaches at 4 TL each. How much money did the grocer in Yeniceabat have in his pocket on Monday evening?



How many peaches were there at the Yeniceabat Grocery Store on Monday evening?



Appendix C
Worksheet for Group B- Yeniceabat Grocery Store (first page)

Monday

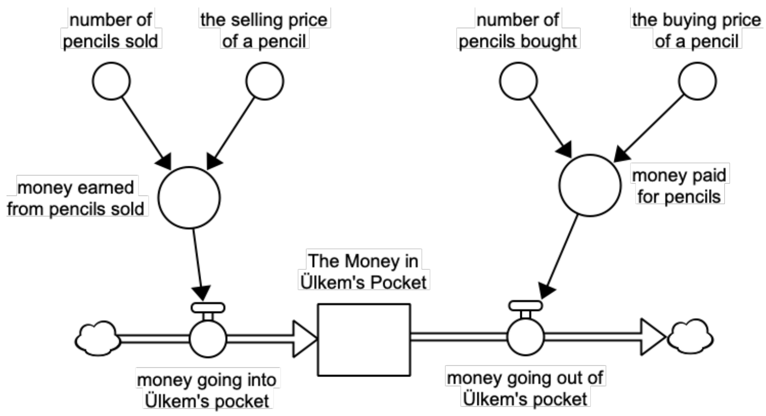
Mr. Umut, the grocer in Yeniceabat Grocery Store, had 50 TL at the beginning of the week. Seeing that he had no peaches left in his shop, Umut Bey bought 10 peaches at 2 TL each on Monday morning. He sold 5 of these peaches at 4 TL each. How much money did the grocer in Yeniceabat have in his pocket on Monday evening?

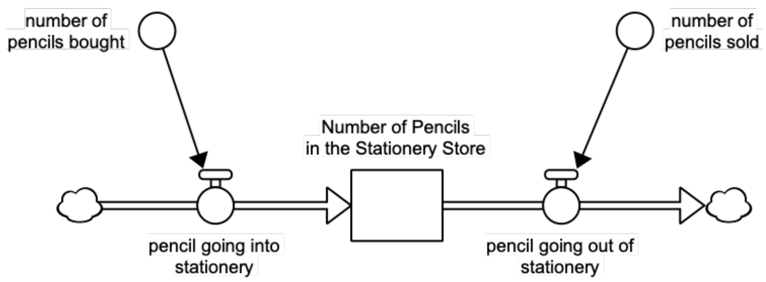
How many peaches were there at the Yeniceabat Grocery Store on Monday evening?

Appendix D
Worksheet for both groups - Ülkem Office Supplies

Monday

Ülkem Office Supplies had 50 TL at the beginning of the week. Seeing that there were 7 pencils left in the store, Mrs. Ülkem bought 4 pencils at 2 TL each. Two pencils were sold that day. She sold each pencil for 4 TL. How much money did Mrs. Ülkem have in her pocket on Monday evening?





Appendix E
R code used in statistical analysis

```
# Code was developed through iterative collaboration with DeepSeek-R1

# Install required packages (if missing)
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("tidyr", quietly = TRUE)) install.packages("tidyr")
if (!requireNamespace("broom", quietly = TRUE)) install.packages("broom")

# Load libraries
library(ggplot2)
library(dplyr)
library(tidyr)
library(broom)

# Read data
data <- read.csv("Data.csv")

# Split data into groups
groupA <- data %>% filter(Group == "A")
groupB <- data %>% filter(Group == "B")

#----------------------------------
# 1. Normality Tests
#----------------------------------
# Explanation of normality testing:
# Before choosing appropriate statistical tests, we need to check if the data follows a normal distribution
# The Shapiro-Wilk test is used here to assess normality:
#   - Null hypothesis (H0): The data is normally distributed
#   - p > 0.05: We cannot reject H0, so we assume the data is normal
#   - p < 0.05: We reject H0, suggesting the data is not normally distributed
#
# We test normality for each combination of group (A/B) and score (Score1/Score2)
# The results of these tests determine which statistical tests we use later

normality_tests <- bind_rows(
  # Test normality for Group A (Score1 and Score2)
  groupA %>% summarise(across(c(Score1, Score2), ~shapiro.test(.)$p.value)) %>% mutate(Group = "A"),

  # Test normality for Group B (Score1 and Score2)
  groupB %>% summarise(across(c(Score1, Score2), ~shapiro.test(.)$p.value)) %>% mutate(Group = "B")
) %>%
  # Convert to long format for easier handling
  pivot_longer(cols = -Group, names_to = "Score", values_to = "p_value") %>%
  # Add a column indicating whether the data is normal or not
  mutate(Normal = ifelse(p_value > 0.05, "Yes", "No"),
         # Add interpretation column for clarity
         Interpretation = ifelse(p_value > 0.05,
                                 "Normally distributed",
                                 "Not normally distributed"))

cat("Normality Test Results:\n")
print(normality_tests, digits = 3)

#----------------------------------
# 2. Statistical Tests
#----------------------------------
# Explanation of test selection:
# The choice of statistical test depends on two main factors:
# 1. Whether the data is paired (within-subject) or unpaired (between-subject)
```

```
# 2. Whether the data follows a normal distribution
#
# For normal data:
#   - Paired data: Paired t-test
#   - Unpaired data: Independent samples t-test (Welch's t-test for unequal variances)
#
# For non-normal data:
#   - Paired data: Wilcoxon signed-rank test
#   - Unpaired data: Mann-Whitney U test (Wilcoxon rank-sum test)
#
# The function below automatically selects the appropriate test based on these criteria

# Function to select appropriate test
run_test <- function(data1, data2, paired = FALSE, normal1, normal2) {
  if (paired) {
    # For paired/within-subject comparisons (e.g., before vs after)
    if (normal1 & normal2) {
      # If both datasets are normally distributed, use paired t-test
      t.test(data1, data2, paired = TRUE)
    } else {
      # If either dataset is not normal, use Wilcoxon signed-rank test (non-parametric alternative)
      # exact=FALSE prevents errors with ties in the data
      wilcox.test(data1, data2, paired = TRUE, exact = FALSE)
    }
  } else {
    # For unpaired/between-subject comparisons (e.g., Group A vs Group B)
    if (normal1 & normal2) {
      # If both datasets are normally distributed, use Welch's t-test (unequal variances)
      # Note: var.equal=FALSE uses Welch's correction, which is generally recommended
      t.test(data1, data2, var.equal = FALSE)
    } else {
      # If either dataset is not normal, use Mann-Whitney U test (non-parametric alternative)
      # exact=FALSE prevents errors with ties in the data
      wilcox.test(data1, data2, exact = FALSE)
    }
  }
}

# Get normality flags from the previous normality tests
# This creates a nested list structure for easy access to normality status
norm_flags <- normality_tests %>%
  mutate(Score = ifelse(Score == "Score1", "Score1", "Score2")) %>%
  split(list(.$Group, .$Score))

# Perform the appropriate statistical tests based on data characteristics:

# 1. Between-groups comparison at Stage 1 (Score1)
# This tests if Group A and Group B had different scores at baseline (Score1)
# Unpaired test: Independent samples t-test if normal, Mann-Whitney if non-normal
stage1_test <- run_test(groupA$Score1, groupB$Score1,
                        normal1 = norm_flags$A.Score1$Normal == "Yes",
                        normal2 = norm_flags$B.Score1$Normal == "Yes")

# 2. Between-groups comparison at Stage 2 (Score2)
# This tests if Group A and Group B had different scores after the intervention (Score2)
# Unpaired test: Independent samples t-test if normal, Mann-Whitney if non-normal
stage2_test <- run_test(groupA$Score2, groupB$Score2,
                        normal1 = norm_flags$A.Score2$Normal == "Yes",
                        normal2 = norm_flags$B.Score2$Normal == "Yes")

# 3. Within-group comparison for Group B (Score1 vs Score2)
# This tests if Group B showed significant change when adopting the new method
```

```
# Paired test: Paired t-test if normal, Wilcoxon signed-rank if non-normal
groupB_paired <- run_test(groupB$Score2, groupB$Score1, paired = TRUE,
                          normal1 = norm_flags$B.Score2$Normal == "Yes",
                          normal2 = norm_flags$B.Score1$Normal == "Yes")

# 4. Within-group comparison for Group A (Score1 vs Score2)
# This tests if Group A showed any significant change despite using the same method
# Paired test: Paired t-test if normal, Wilcoxon signed-rank if non-normal
groupA_paired <- run_test(groupA$Score2, groupA$Score1, paired = TRUE,
                          normal1 = norm_flags$A.Score2$Normal == "Yes",
                          normal2 = norm_flags$A.Score1$Normal == "Yes")

# Create a comprehensive results table
# The broom::tidy() function converts test results into dataframes for easier reporting
# We combine all test results into a single table for better comparison
results_table <- bind_rows(
  # Between-group comparisons
  tidy(stage1_test) %>% mutate(Test = "Stage 1 (A vs B)",
                               Type = "Between-groups",
                               Description = "Tests if groups differed at baseline"),

  tidy(stage2_test) %>% mutate(Test = "Stage 2 (A vs B)",
                               Type = "Between-groups",
                               Description = "Tests if groups differed after intervention"),

  # Within-group comparisons
  tidy(groupB_paired) %>% mutate(Test = "Group B (Score2 vs Score1)",
                                 Type = "Within-group",
                                 Description = "Tests if Group B improved with new method"),

  tidy(groupA_paired) %>% mutate(Test = "Group A (Score2 vs Score1)",
                                 Type = "Within-group",
                                 Description = "Tests if Group A changed over time")
) %>%
  # Select the most important columns and format numbers
  select(Test, Type, Description, estimate, statistic, p.value, method) %>%
  mutate(across(where(is.numeric), ~round(., 3)),
       # Add significance indicators
       Significant = ifelse(p.value < 0.05, "Yes", "No"))

cat("\nStatistical Test Results:\n")
print(results_table)

#---------------------------------
# 3. Visualization
#---------------------------------
# Prepare data
data_long <- data %>%
  pivot_longer(
    cols = c(Score1, Score2),
    names_to = "Stage",
    values_to = "Score"
  ) %>%
  mutate(
    Stage = case_when(
      Stage == "Score1" ~ "Stage 1",
      Stage == "Score2" ~ "Stage 2"
    )
  )

delta_data <- data %>% mutate(Delta = Score2 - Score1)
```

```
# Plot 1: Between-group comparisons
p1 <- ggplot(data_long, aes(x = Group, y = Score, fill = Stage)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.5) +
  stat_summary(fun = mean, geom = "point", shape = 19, size = 3, color = "white") +
  facet_wrap(~Stage) +
  labs(title = "Between-Group Performance Comparison",
       subtitle = "Stage 1: Different methods | Stage 2: Both use new method") +
  theme_minimal()

ggsave("plot1_between_groups.png", p1, width = 8, height = 5)

# Plot 2: Within-group changes
p2 <- ggplot(delta_data, aes(x = Group, y = Delta, fill = Group)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Within-Group Improvement (Score2 - Score1)",
       y = "Improvement") +
  theme_minimal()

ggsave("plot2_within_groups.png", p2, width = 6, height = 5)

# Plot 3: Density distributions
p3 <- ggplot(data_long, aes(x = Score, fill = Group)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Stage, ncol = 1) +
  labs(title = "Score Distribution Shifts Between Stages") +
  theme_minimal()

ggsave("plot3_distributions.png", p3, width = 6, height = 8)

#----------------------------------
# 4. Conclusion and Interpretation of Test Results
#----------------------------------
# This section summarizes the key findings from the statistical analysis
# The interpretation is based on the p-values from the statistical tests and visualization patterns

cat("\nKey Conclusions and Statistical Interpretation:\n",
    "1. Between-group comparison (Stage 1): ",
    ifelse(stage1_test$p.value < 0.05,
           "Significant difference detected between Group A and B at baseline (p < 0.05). This indicates the groups started with different performance levels.",
           "No significant difference between groups at baseline (p > 0.05)."),
    "\n",
    "2. Between-group comparison (Stage 2): ",
    ifelse(stage2_test$p.value < 0.05,
           "Groups remained significantly different after the intervention (p < 0.05).",
           "No significant difference between groups after the intervention (p > 0.05). The gap between groups appears to have closed."),
    "\n",
    "3. Within-group change (Group B): ",
    ifelse(groupB_paired$p.value < 0.05,
           "Group B showed significant improvement when adopting the new method (p < 0.05). This suggests the new method was effective for Group B.",
           "Group B showed no significant change after adopting the new method (p > 0.05)."),
    "\n",
    "4. Within-group change (Group A): ",
    ifelse(groupA_paired$p.value < 0.05,
           "Group A showed significant change over time (p < 0.05).",
           "Group A maintained consistent performance over time (p > 0.05). This is expected as Group A used the same method throughout."),
```

```
  "\n",
  "5. Test selection: ",
  if(all(normality_tests$Normal == "Yes")) {
    "All data were normally distributed, so parametric tests (t-tests) were used."
  } else {
    paste("Some data were not normally distributed (",
          sum(normality_tests$Normal == "No"), " of ", nrow(normality_tests),
          " distributions), so appropriate non-parametric alternatives were used where needed.")
  },
  sep = "")
```

Session information (sessionInfo() output)

```
R version 4.3.1 (2023-06-16)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS 15.5

Matrix products: default
BLAS:
/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLA
S.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK version
3.11.0

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Europe/Istanbul
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ggrepel_0.9.6      RColorBrewer_1.1-3 kableExtra_1.4.0   psych_2.3.6        corrplot_0.95      ggpubr_0.6.0
knitr_1.43
 [8] gridExtra_2.3      svglite_2.1.3      broom_1.0.5        lubridate_1.9.2    forcats_1.0.0      stringr_1.5.0      dplyr_1.1.2
[15] purrr_1.0.1        readr_2.1.4        tidyr_1.3.0        tibble_3.2.1       ggplot2_3.4.2      tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] gtable_0.3.3      xfun_0.39         rstatix_0.7.2     lattice_0.21-8    tzdb_0.4.0        vctrs_0.6.3       tools_4.3.1
generics_0.1.3
 [9] parallel_4.3.1    fansi_1.0.4       highr_0.10        pkgconfig_2.0.3   Matrix_1.5-4.1    lifecycle_1.0.3   compiler_4.3.1
farver_2.1.1
[17] textshaping_0.3.6 munsell_0.5.0     mnormt_2.1.1      carData_3.0-5     htmltools_0.5.5   Formula_1.2-5
pillar_1.9.0      car_3.1-3
[25] crayon_1.5.2      abind_1.4-8       nlme_3.1-162      tidyselect_1.2.0  digest_0.6.33     stringi_1.7.12    labeling_0.4.2
splines_4.3.1
[33] cowplot_1.1.3     fastmap_1.1.1     grid_4.3.1        colorspace_2.1-0  cli_3.6.1         magrittr_2.0.3    utf8_1.2.3
withr_2.5.0
[41] scales_1.2.1      backports_1.4.1   bit64_4.0.5       timechange_0.2.0  rmarkdown_2.23    bit_4.0.5         ggsignif_0.6.4
ragg_1.2.5
[49] hms_1.1.3         evaluate_0.21     viridisLite_0.4.2 mgcv_1.8-42       rlang_1.1.1       Rcpp_1.0.11       glue_1.6.2
xml2_1.3.5
[57] rstudioapi_0.15.0 vroom_1.6.3       R6_2.5.1          systemfonts_1.0.4
```