Exploratory Data Analysis Report - The Residuals

Bhagatinder Longia, Musab Muhammad, Muntasir Munem, Shea Munson, Chloe Syriac

February 28, 2025

Summary Statistic Tables

In our exploratory data analysis, we created the following tables to summarize key information about the quantitative and qualitative variables involved in our study.

Table 1: Summary of Problem Solving Test Scores

	Mean	Median	SD	IQR
Value	93.5	96	6.9	6.8

The above table contains statistics for the scores obtained by the participants of our study on the problem solving test. The high mean and median indicate that participants generally scored very high on the test. Furthermore, the standard deviation and the IQR are nearly identical, indicating that the scores are tightly clustered around the mean. This implies that there is relatively little overall variability in the scores.

The tables below show how we employed a fully balanced design in our data collection, ensuring that the participants are evenly distributed across the different age groups and drink types.

Table 2: Summary of Age Groups

	18-35	36-54	55+
Count	50	50	50
Percentage	33.3	33.3	33.3

Table 3: Summary of Types of Drinks

	Water	Coffee	Coffee Decaffeinated	Energy Drink	Energy Drink Caffeine-Free
Count	30	30	30	30	30
Percentage	20	20	20	20	20

Within each age group, there were 10 observations per drink type (so a total of 50 observations since there are 5 drink types), which allowed us to use age as a blocking factor while assessing the impact of the drink type on the problem solving scores of the participants.

Plots

We created the following box-plots to compare the distribution of the problem solving scores by the treatment (drink type) given to the participants and the age group of the participants.

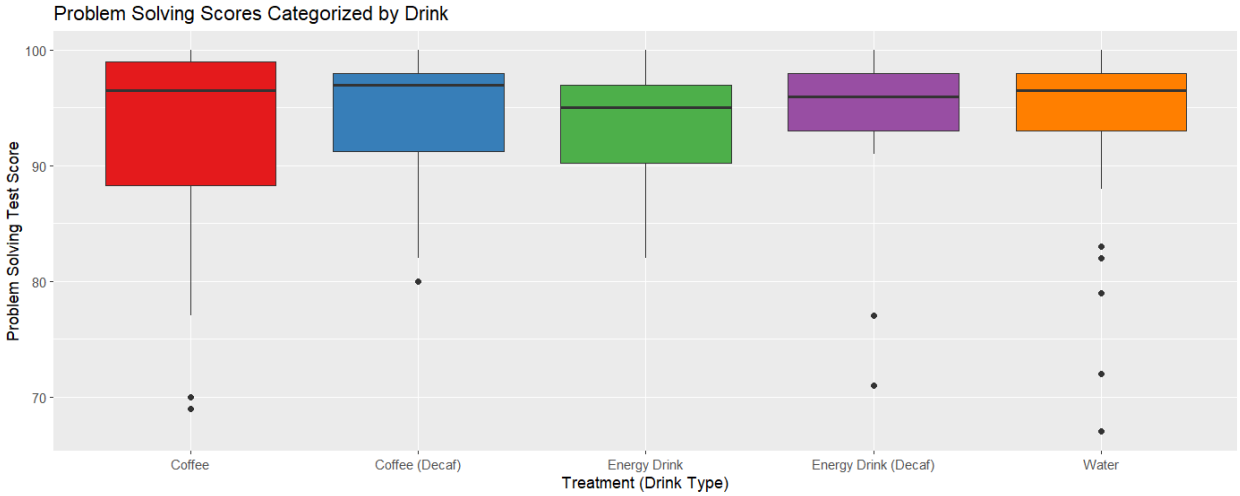

Figure 1: Scores Categorized by Drink Type

We can see from Figure 1 that the median score is pretty close among all the participants regardless of the treatment given to them. In terms of variability, we see that Coffee (Caffenated) has the most variability in the scores as it has relatively larger box (bigger IQR) and longer whiskers compared to the other drinks. The scores in Energy drink (decaffenated) seem to be more centered around the median as it has the smallest box (smaller IQR) and short whiskers. Most drinks have one to two outliers, however water has the most outliers out of all the drinks.

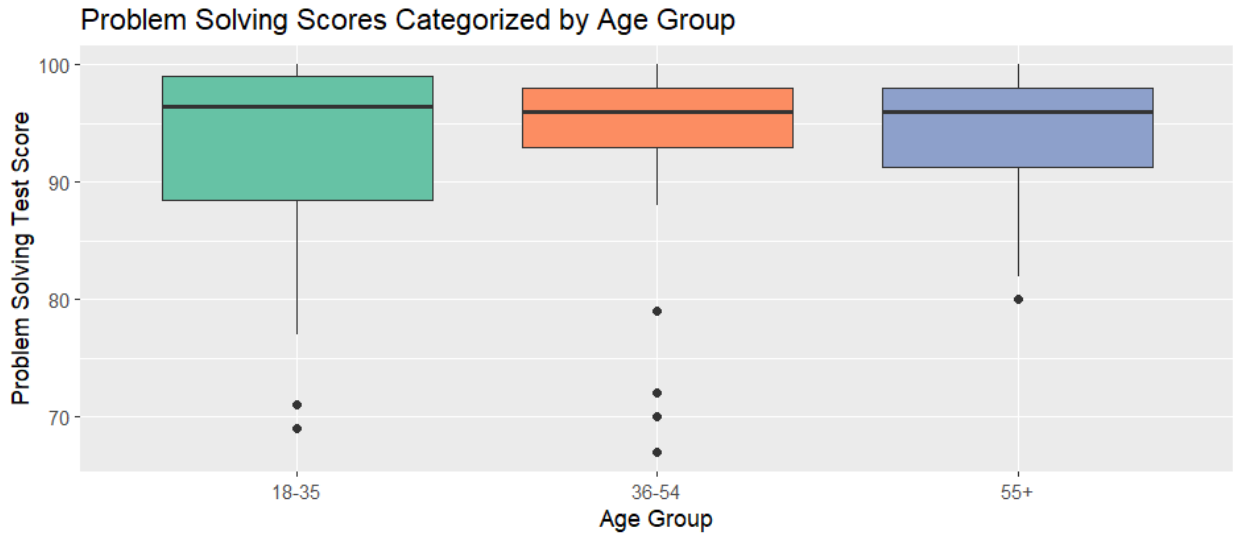

Figure 2: Scores Categorized by Age Group

We can see from Figure 2 that the median score is pretty close among all the participants regardless of the age group they belong to. In terms of variability, we see that the scores of the 18-35 age group have the most variability in the scores. The age group of 36-54 has the least variability but also the most outliers out of all the age groups.

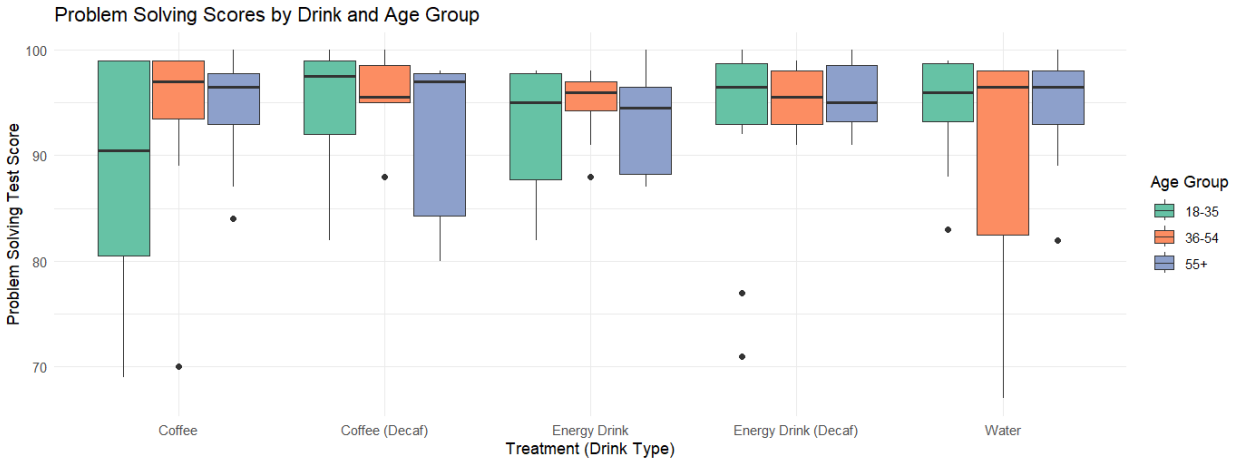

Figure 3: Scores Categorized by both Drink and Age Group

Finally from Figure 3, as we saw before, we can see that the medians are pretty close. There is high variability in the scores of participants who were given Coffee (caffenated) particularly in the 18-35 age group. Participants who were given water also have high variability in the 36-54 age group.

Assumptions

For our tests we decided to do a one-factor test (with just the treatments as a factor) and a two-factor test (with the blocking factor of age also added as a factor). For both tests, we found that assumptions of normality and homogeneity of variance were heavily violated.

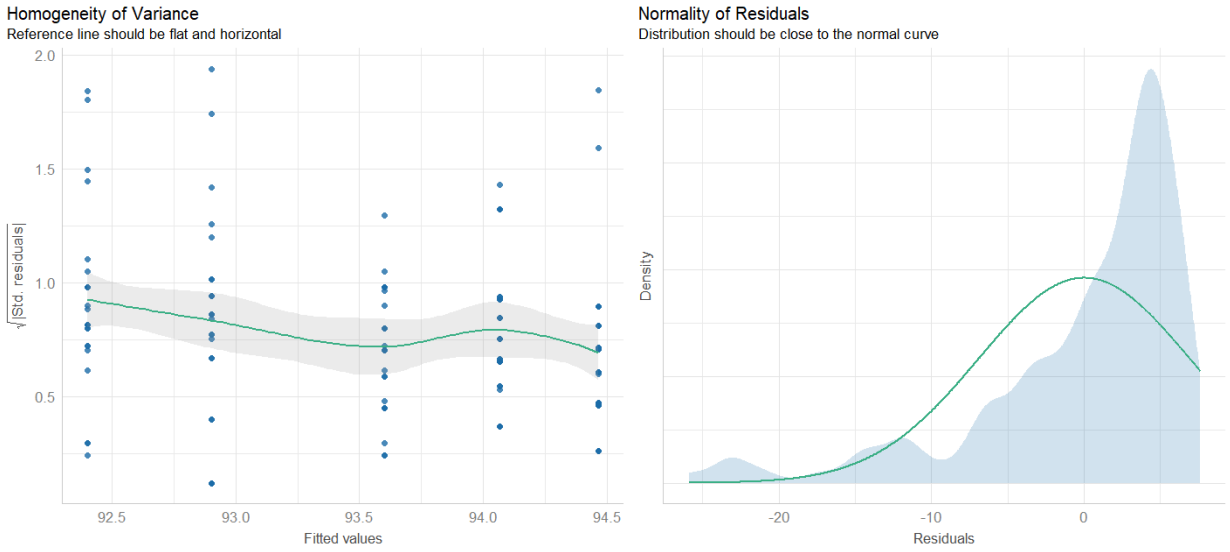

Figure 4: One-factor Anova Test Assumption Checks

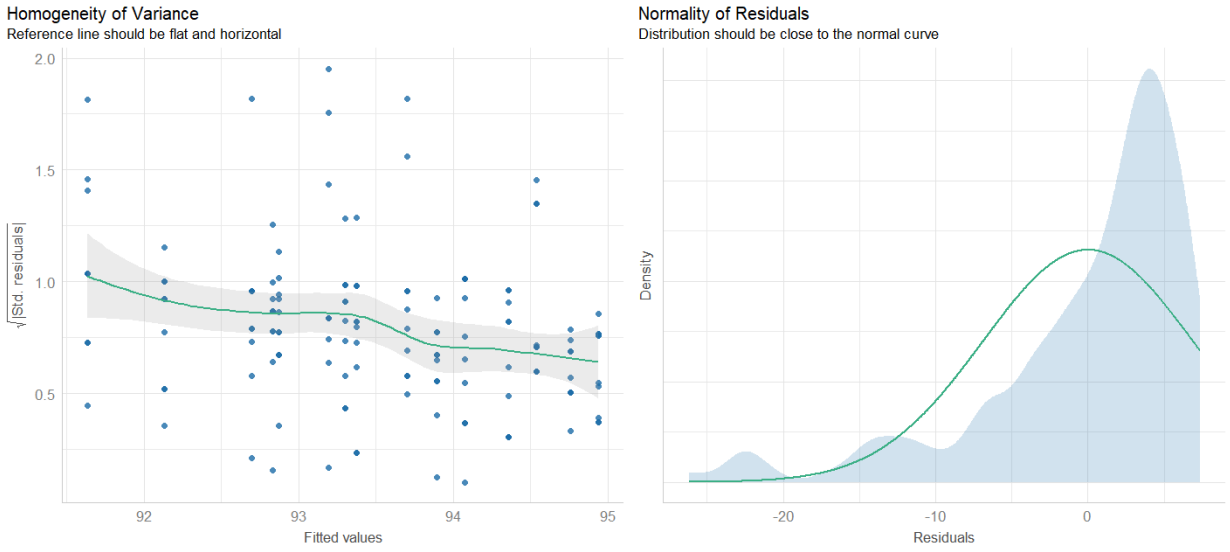

Figure 5: Two-factor Anova Test Assumption Checks

From the homogenity of variance graphs on the left side of Figure 4 and Figure 5, we can see that the residual points do not follow the reference line, showing that the data does not

have equal variance, similarly, from the check_heteroscedasticity function called on both our one and two factor ANOVA models, we find that this assumption is heavily violated (both tests had p-values less than 0.003).

From the normality of residuals graphs on the right side of Figure 4 and Figure 5, we can see that the distribution is not close to the normal curve showing that the data is not normally distributed, similarly, from the check_normality function called on both our one and two factor ANOVA models, we find that this assumption is also heavily violated (both tests had p-values less than 0.001).

Appendix

Overleaf: https://www.overleaf.com/project/67bff89e59232b6befedfe09

```r
# STA305 Project R Code

# List of required packages
packages <- c("performance", "tidyverse", "knitr", "patchwork", "see")

# Install any packages that aren't already installed
for(pkg in packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}

# Load the packages
library(performance)
library(tidyverse)
library(knitr)
library(patchwork)

# Load the data
data <- read.csv("cleaned_data.csv", header = TRUE)

# Check the first few rows and structure
head(data)
str(data)

# Generating summary table for Problem Solving Scores
summary_table <- data.frame(Value = c(mean(data$score),
                                      median(data$score),
                                      sd(data$score),
                                      IQR(data$score)))
kable(t(summary_table),
      col.names = c("Mean", "Median", "SD", "IQR"),
      caption = "Summary of Problem Solving Test Scores",
      align = "c")

# Generating summary table for Summary of Age Group of participants
data$age_group <- factor(data$age_group,
                         levels = c("Y", "M", "O"),
                         labels = c("18-35", "36-54", "55+"))
age_dist <- table(data$age_group)
age_percent <- prop.table(age_dist) * 100
age_summary <- rbind(
  Count = as.numeric(age_dist),
  Percentage = round(as.numeric(age_percent), 1)
)
colnames(age_summary) <- names(age_dist)
kable(age_summary, caption = "Summary of Age Groups of participants")

# Generating summary table for Summary of Treatments
data$treatment <- factor(data$treatment,
                         levels = c("C", "CD", "E", "ED", "W"),
                         labels = c("Coffee", "Coffee (Decaf)", "Energy Drink",
                                    "Energy Drink (Decaf)", "Water"))
treatment_dist <- table(data$treatment)
```

```r
treatment_percent <- prop.table(treatment_dist) * 100
treatment_summary <- rbind(
  Count = as.numeric(treatment_dist),
  Percentage = round(as.numeric(treatment_percent), 1)
)
colnames(treatment_summary) <- names(treatment_dist)
kable(treatment_summary, caption = "Summary of Types of Drink given")

# Generating box plot for scores in relation to the treatments
tr <- ggplot(data, aes(x = treatment, y = score, fill = treatment)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("Coffee", "Coffee (Decaf)", "Energy Drink",
                              "Energy Drink (Decaf)", "Water")) +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Problem Solving Scores Categorized by Drink",
       x = "Treatment (Drink Type)", y = "Problem Solving Test Score") +
  theme(legend.position = "none")

# Generating box plot for scores in relation to the age of participants
ag <- ggplot(data, aes(x = age_group, y = score, fill = age_group)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Set2") +
  scale_x_discrete(labels = c("18-35", "36-54", "55+")) +
  labs(title = "Problem Solving Scores Categorized by Age Group",
       x = "Age Group", y = "Problem Solving Test Score") +
  theme(legend.position = "none")

# Print the individual plots
print(tr)
print(ag)

# Generating combined box plot for scores by drink and age group
combined_plot <- ggplot(data, aes(x = treatment, y = score, fill = age_group)) +
  geom_boxplot(position = position_dodge(width = 0.8)) +
  scale_x_discrete(labels = c("Coffee", "Coffee (Decaf)", "Energy Drink",
                              "Energy Drink (Decaf)", "Water")) +
  labs(title = "Problem Solving Scores by Drink and Age Group",
       x = "Treatment (Drink Type)",
       y = "Problem Solving Test Score",
       fill = "Age Group") +
  scale_fill_brewer(palette = "Set2",
                    breaks = c("18-35", "36-54", "55+"),
                    labels = c("18-35", "36-54", "55+")) +
  theme_minimal()

print(combined_plot)


# Generating the graphs to check for normality and homogeneity assumptions
# for the one-factor ANOVA model and the two-factor ANOVA model
one_factor_model <- aov(data$score ~ data$treatment, data = data)
check_model(one_factor_model, check = c("normality", "homogeneity"))

two_factor_model <- aov(data$score ~ data$treatment + data$age_group, data = data)
check_model(two_factor_model, check = c("normality", "homogeneity"))

# Print results of assumption checks:
# 1. Check normality of residuals for the one-factor model
```

```r
normality_results1 <- check_normality(one_factor_model)
print(normality_results1)

# 2. Check homogeneity of variances for the one-factor model
hetero_results1 <- check_heteroscedasticity(one_factor_model)
print(hetero_results1)

# 1. Check normality of residuals for the one-factor model again
normality_results2 <- check_normality(one_factor_model)
print(normality_results2)

# 2. Check homogeneity of variances for the two-factor model
hetero_results2 <- check_heteroscedasticity(two_factor_model)
print(hetero_results2)
```
