

# Untitled

Simran Gill

2024-11-27

## Loading data

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.4.1
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.4.1
```

```
library(stargazer)
```

```
##
```

```
## Please cite as:
```

```
## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
```

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
```

```
##
```

```
## combine
```

```
## The following objects are masked from 'package:data.table':
```

```
##
```

```
## between, first, last
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## filter, lag
```

```
## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.1
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
## recode
```

```
control <- read.csv("Coffee Survey Control (Responses) - Form Responses 1.csv")
treatment <- read.csv("Coffee Survey Group II (Responses) - Form Responses 1.csv")
```

```
# rename column names for control
colnames(control) <- c('timestamp', 'good_and_gather_score', 'chameleon_score', 'age', 'gender', 'how_of')
# rename column names for treatment
colnames(treatment) <- c('timestamp', 'name', 'good_and_gather_score', 'chameleon_score', 'age', 'gender')
# reorder column names for treatment
treatment <- treatment[, c('timestamp', 'good_and_gather_score', 'chameleon_score', 'age', 'gender', 'how_of')]

control$treatment <- 0
control$age <- as.integer(control$age)
```

```
## Warning: NAs introduced by coercion
```

```
# re-labeling gender
control <- control %>%
  mutate(gender = case_when(
    gender == "F" ~ "Female",
    gender == "M" ~ "Male",
    TRUE ~ "Unknown"
  ))

treatment$treatment <- 1
treatment$age <- as.integer(treatment$age)

# re-labeling gender
treatment <- treatment %>%
  mutate(gender = case_when(
    gender == "F" ~ "Female",
    gender == "M" ~ "Male",
    TRUE ~ "Unknown"
  ))
```

```

))

d <- rbind(control, treatment)

cat("Number of Rows before cleaning:",nrow(d))

## Number of Rows before cleaning: 92

# removing rows where age is null
d <- d %>%
  filter(!is.na(age))

# creating age groups
d$age_group <- cut(d$age,
  breaks = c(0, 20, 30, 40, 50, Inf),
  labels = c("Under 20", "20-30", "31-40", "41-50", "Over 50"),
  right = FALSE)

# Convert how_often_drink_coffee to integer by factoring
d$how_often_drink_coffee <- factor(d$how_often_drink_coffee,
  levels = c("Never",
    "Occasionally (up to 1 time a week)",
    "Sometimes (a few times a week)",
    "Often (almost every day)",
    "Every day"))

d$chameleon_awareness_flag <- ifelse(d$chameleon_awareness == "No", 0, 1)
d$good_and_gather_awareness_flag <- ifelse(d$good_and_gather_awareness == "No", 0, 1)

cat("\nNumber of Rows after cleaning:",nrow(d))

##
## Number of Rows after cleaning: 90

str(d)

```

```

## 'data.frame': 90 obs. of 16 variables:
## $ timestamp : chr "11/11/2024 9:36:42" "11/11/2024 9:38:27" "11/11/2024 9:41:50" ...
## $ good_and_gather_score : int 3 3 1 3 1 5 4 5 3 5 ...
## $ chameleon_score : int 5 5 5 5 3 4 2 3 5 5 ...
## $ age : int 34 21 27 23 24 35 24 24 23 24 ...
## $ gender : chr "Male" "Male" "Female" "Female" ...
## $ how_often_drink_coffee : Factor w/ 5 levels "Never","Occasionally (up to 1 time a week)",...
## $ hot_or_cold : chr "Hot Coffee" "Hot Coffee" "Cold Coffee" "Hot Coffee" ...
## $ sweet_or_not_sweet : chr "Not Sweet" "Sweet" "Not Sweet" "Not Sweet" ...
## $ good_and_gather_awareness : chr "No" "Yes, Neutral" "Yes, Positive" "Yes, Positive" ...
## $ chameleon_awareness : chr "No" "No" "Yes, Positive" "No" ...
## $ medical_condition : chr "No" "No" "No" "No" ...
## $ name : chr "Kavin" "Arya Desai" "Liz Ren" "Halal Biviji" ...
## $ treatment : num 0 0 0 0 0 0 0 0 0 0 ...
## $ age_group : Factor w/ 5 levels "Under 20","20-30",...: 3 2 2 2 2 3 2 2 2 2 ...

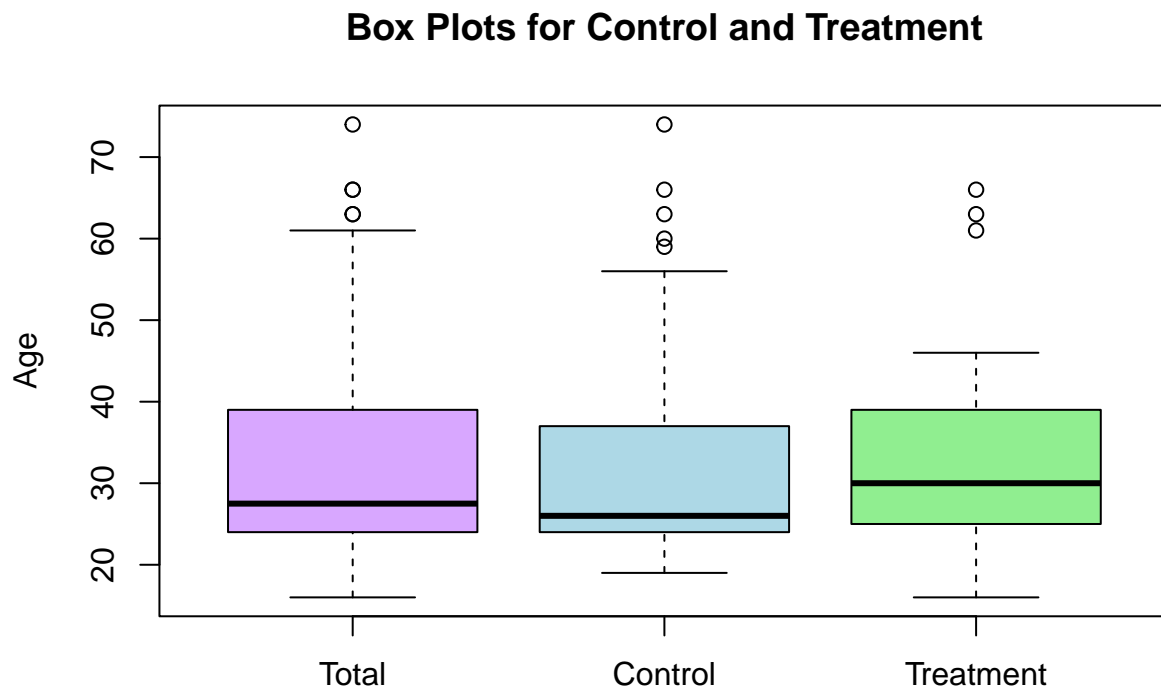
```

```
## $ chameleon_awareness_flag      : num  0 0 1 0 1 1 0 0 1 0 ...
## $ good_and_gather_awareness_flag: num  0 1 1 1 0 1 0 0 1 1 ...
```

## Exploratory Data Analysis

```
# box plot for age by treatment and control

boxplot(d$age, control$age, treatment$age,
        names = c("Total", "Control", "Treatment"),
        main = "Box Plots for Control and Treatment",
        ylab = "Age",
        col = c("#D8A7FF", "lightblue", "lightgreen"),
        border = "black")
```



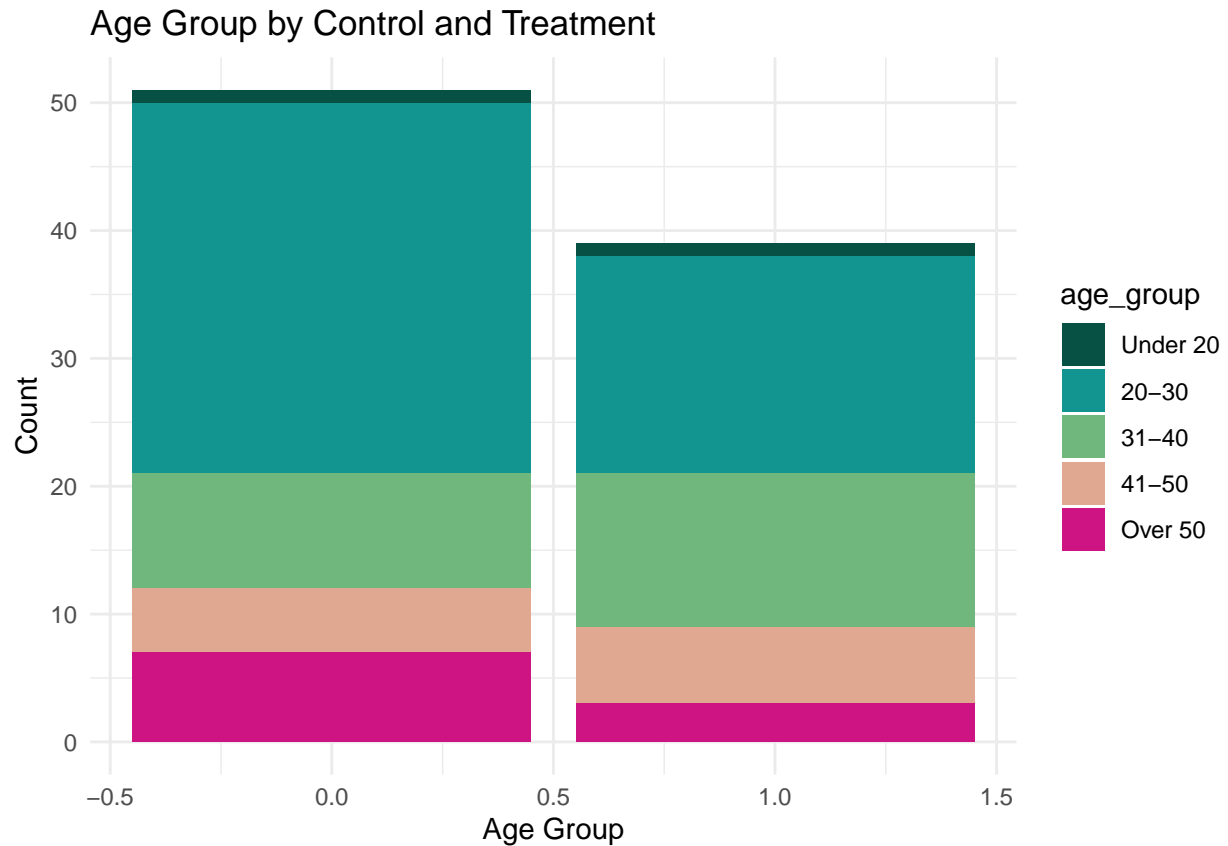
```
cat("Number of Rows for Treatment Group:", sum(d$treatment == 1))
```

```
## Number of Rows for Treatment Group: 39
```

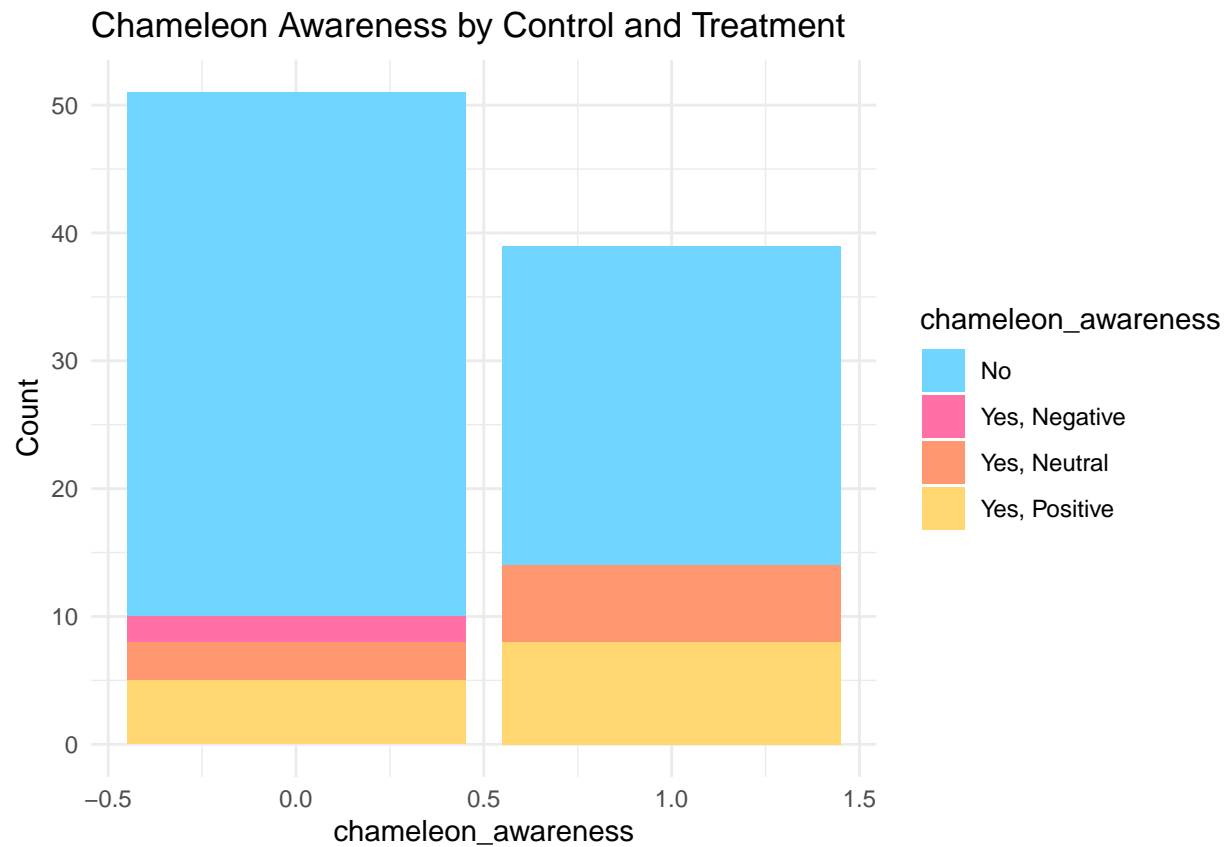
```
cat("\nNumber of Rows for Control Group:", sum(d$treatment == 0))
```

```
##
## Number of Rows for Control Group: 51
```

```
ggplot(d, aes(x = treatment, fill = age_group)) +
  geom_bar(position = "stack") +
  labs(title = "Age Group by Control and Treatment", x = "Age Group", y = "Count") +
  scale_fill_manual(values = c("#065143", "#129490", "#70B77E", "#E0A890", "#CE1483")) +
  theme_minimal()
```

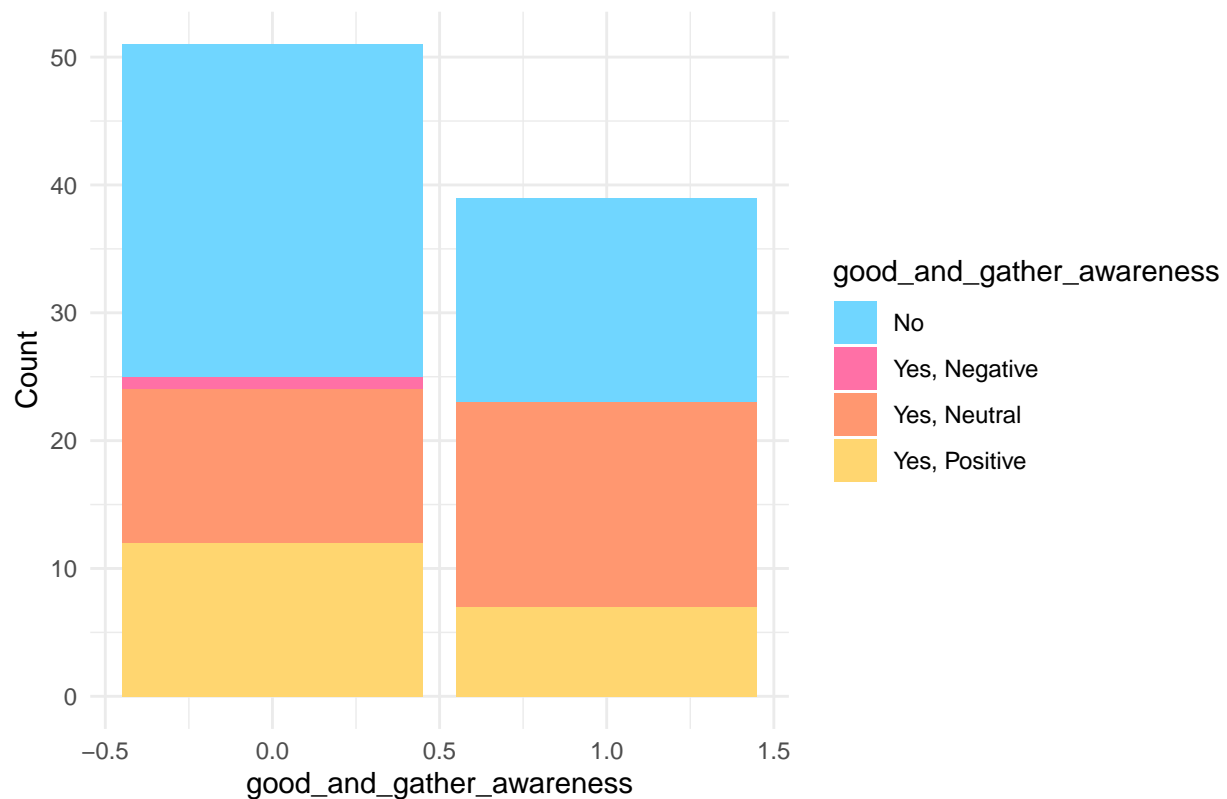


```
ggplot(d, aes(x = treatment, fill = chameleon_awareness)) +
  geom_bar(position = "stack") +
  labs(title = "Chameleon Awareness by Control and Treatment", x = "chameleon_awareness", y = "Count") +
  scale_fill_manual(values = c("#70D6FF", "#FF70A6", "#FF9770", "#FFD670")) +
  theme_minimal()
```



```
ggplot(d, aes(x = treatment, fill = good_and_gather_awareness)) +
  geom_bar(position = "stack") +
  labs(title = "Good&Gather Awareness by Control and Treatment", x = "good_and_gather_awareness", y = "Count") +
  scale_fill_manual(values = c("#70D6FF", "#FF70A6", "#FF9770", "#FFD670")) +
  theme_minimal()
```

## Good&Gather Awareness by Control and Treatment



```
### Control Group Gender ###
control_gender_counts <- control %>%
  group_by(gender) %>%
  tally()

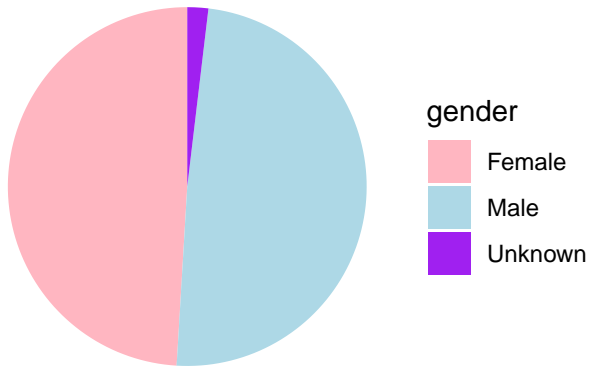
control_pie <- ggplot(control_gender_counts, aes(x = "", y = n, fill = gender)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Gender Distribution for Control Group") +
  scale_fill_manual(values = c("lightpink", "lightblue", "purple")) +
  theme_void()

### Treatment Group Gender ###
treatment_gender_counts <- treatment %>%
  group_by(gender) %>%
  tally()

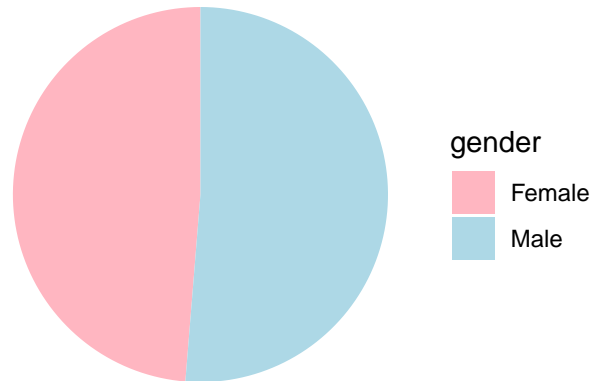
treatment_pie <- ggplot(treatment_gender_counts, aes(x = "", y = n, fill = gender)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Gender Distribution for Treatment Group") +
  scale_fill_manual(values = c("lightpink", "lightblue", "purple")) +
  theme_void()

grid.arrange(control_pie, treatment_pie, ncol = 2)
```

Gender Distribution for Control Group



Gender Distribution for Treatment Group



## Simple Average Treatment Effect

```
ate_good_and_gather <- mean(d$good_and_gather_score[d$treatment == 1], na.rm = TRUE) -
                        mean(d$good_and_gather_score[d$treatment == 0], na.rm = TRUE)
cat("ATE Good & Gather:", ate_good_and_gather)
```

```
## ATE Good & Gather: -0.4434389
```

```
ate_chameleon <- mean(d$chameleon_score[d$treatment == 1], na.rm = TRUE) -
                  mean(d$chameleon_score[d$treatment == 0], na.rm = TRUE)
cat("\nATE Chameleon:", ate_chameleon)
```

```
##
```

```
## ATE Chameleon: 0.2865762
```

## Average Treatment Effect using Linear Regression

```
# Basic Linear regression to estimate ATE
model_gg <- lm(good_and_gather_score ~ treatment, data=d)
ate_regression <- coef(model_gg)["treatment"]
print(ate_regression)
```



```
## treatment
## -0.4434389
```

```
summary(model_gg)
```

```
##
## Call:
## lm(formula = good_and_gather_score ~ treatment, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05882 -1.05882 -0.05882  0.94118  2.94118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.0588      0.1992  20.377  <2e-16 ***
## treatment    -0.4434      0.3026  -1.466   0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.422 on 88 degrees of freedom
## Multiple R-squared:  0.02383,    Adjusted R-squared:  0.01273
## F-statistic: 2.148 on 1 and 88 DF,  p-value: 0.1463
```

```
# Basic Linear regression to estimate ATE
model_c <- lm(chameleon_score ~ treatment, data=d)
ate_regression <- coef(model_c)["treatment"]
print(ate_regression)
```

```
## treatment
## 0.2865762
```

```
summary(model_c)
```

```
##
## Call:
## lm(formula = chameleon_score ~ treatment, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0513 -1.5863  0.2353  1.2353  2.2353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7647      0.2169  17.36  <2e-16 ***
## treatment     0.2866      0.3294   0.87   0.387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.549 on 88 degrees of freedom
## Multiple R-squared:  0.008525,    Adjusted R-squared:  -0.002741
## F-statistic: 0.7567 on 1 and 88 DF,  p-value: 0.3867
```

## ATE Adjusted for Covariates

```
model_gg_covariates <- lm(good_and_gather_score ~ treatment + log(age) + gender + chameleon_awareness ,
ate_with_covariates <- coef(model_gg_covariates)["treatment"]
print(ate_with_covariates)
```

```
## treatment
## -0.5684904
```

```
summary(model_gg_covariates)
```

```
##
## Call:
## lm(formula = good_and_gather_score ~ treatment + log(age) + gender +
##     chameleon_awareness, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10730 -0.82418 -0.00309  0.84285  2.56540
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.5236     1.5124  -0.346  0.73005
## treatment      -0.5685     0.2870  -1.981  0.05094 .
## log(age)        1.3032     0.4280   3.045  0.00312 **
## genderMale       0.3850     0.2764   1.393  0.16736
## chameleon_awarenessYes, Negative -1.8873     0.9450  -1.997  0.04908 *
## chameleon_awarenessYes, Neutral  0.9324     0.4750   1.963  0.05298 .
## chameleon_awarenessYes, Positive -0.6172     0.4014  -1.538  0.12792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.301 on 83 degrees of freedom
## Multiple R-squared:  0.2293, Adjusted R-squared:  0.1736
## F-statistic: 4.116 on 6 and 83 DF,  p-value: 0.00115
```

```
model_gg_covariates_v2 <- lm(good_and_gather_score ~ treatment + gender + log(age) + chameleon_awareness
anova(model_gg_covariates , model_gg_covariates_v2)
```

```
## Analysis of Variance Table
##
## Model 1: good_and_gather_score ~ treatment + log(age) + gender + chameleon_awareness
## Model 2: good_and_gather_score ~ treatment + gender + log(age) + chameleon_awareness +
##     good_and_gather_awareness
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      83 140.57
## 2      80 136.00  3    4.5703 0.8961 0.447
```

**Interpretation** We tested multiple covariates to see if we can improve the regression model for Good&Gather Score. The main covariates we see has a positive impact is how a participant views the Chameleon brand and age group.

When it comes to age, participants in the age group 30 - 39 and 40 - 49 are likely to rate Good & Gather higher after the brand is revealed. Because these two variables have some significant, age group does play a part in how a participant rates the coffee after treatment is provided.

When it comes to the Chameleon, even though the participants has a negative view of Chameleon coffee as a brand, they are still likely to score Good & Gather -1.8285 after treatment is provided. The p-value for Chameleon awareness is 0.0561, which means this variable is marginally significant.

We also wanted to test if adding Good&Gather brand awareness as a variable to model has an significant effect to the model. From the ANOVA test we can see that the p-value is 0.5600 which is greater than 0.05. This indicated Good&Gather brand awareness has no statistically significant impact on scoring the coffee.

```
stargazer(model_gg_covariates, model_gg_covariates_v2,
  type = "text",    # Use "html" for HTML output or "latex" for LaTeX
  title = "Regression Results for Good and Gather Score",
  covariate.labels = c("Treatment", "log(Age)", "Gender", "Chameleon Awareness", "Good and Gather Awareness"),
  star.cutoffs = c(0.10, 0.05, 0.01, 0.001),    # Significance stars
  out = "regression_table.txt") # Optional: Save output to a text file
```

```
##
## Regression Results for Good and Gather Score
## =====
##                               Dependent variable:
##                               -----
##                               good_and_gather_score
##                               (1)                (2)
## -----
```

## Treatment	-0.568*	-0.513*
##	(0.287)	(0.294)
## log(Age)	1.303***	1.130**
##	(0.428)	(0.442)
## Gender	0.385	0.254
##	(0.276)	(0.289)
## Chameleon Awareness	-1.887**	-2.097**
##	(0.945)	(0.960)
## Good and Gather Awareness	0.932*	0.951*
##	(0.475)	(0.493)
## chameleon_awarenessYes, Positive	-0.617	-0.526
##	(0.401)	(0.421)
## good_and_gather_awarenessYes, Negative		-0.856
##		(1.350)
## good_and_gather_awarenessYes, Neutral		-0.543
##		(0.350)
##		

```
## good_and_gather_awarenessYes, Positive -0.147
## (0.391)
##
## Constant -0.524 0.314
## (1.512) (1.605)
##
## -----
## Observations 90 90
## R2 0.229 0.254
## Adjusted R2 0.174 0.170
## Residual Std. Error 1.301 (df = 83) 1.304 (df = 80)
## F Statistic 4.116*** (df = 6; 83) 3.032*** (df = 9; 80)
## =====
## Note: *p<0.1; **p<0.05; ***p<0.01
```

```
c('timestamp', 'good_and_gather_score', 'chameleon_score', 'age', 'gender', 'how_often_drink_coffee',
'hot_or_cold', 'sweet_or_not_sweet', 'good_and_gather_awareness', 'chameleon_awareness', 'medical_condition', 'name')
```

```
model_c_covariates <- lm(chameleon_score ~ treatment + log(age) + gender + chameleon_awareness_flag + good_and_gather_awareness_flag, data = d)
ate_with_covariates <- coef(model_c_covariates)["treatment"]
print(ate_with_covariates)
```

```
## treatment
## -0.01728814
```

```
summary(model_c_covariates)
```

```
##
## Call:
## lm(formula = chameleon_score ~ treatment + log(age) + gender +
##      chameleon_awareness_flag + good_and_gather_awareness_flag,
##      data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2778 -0.9554  0.0911  1.0778  2.7193
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.72728    1.64566  -1.050  0.296916
## treatment     -0.01729    0.29862  -0.058  0.953970
## log(age)       1.37784    0.45577   3.023  0.003317 **
## genderMale     0.41675    0.30145   1.383  0.170482
## chameleon_awareness_flag 1.30106    0.34630   3.757  0.000316 ***
## good_and_gather_awareness_flag 0.61820    0.31401   1.969  0.052276 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.375 on 84 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.21
## F-statistic: 5.731 on 5 and 84 DF, p-value: 0.0001344
```

```
model_c_covariates_v2 <- lm(chameleon_score ~ treatment + log(age) + gender + chameleon_awareness_flag,
anova(model_c_covariates , model_c_covariates_v2)
```

```
## Analysis of Variance Table
##
## Model 1: chameleon_score ~ treatment + log(age) + gender + chameleon_awareness_flag +
##   good_and_gather_awareness_flag
## Model 2: chameleon_score ~ treatment + log(age) + gender + chameleon_awareness_flag
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      84 158.74
## 2      85 166.06 -1    -7.3248 3.8761 0.05228 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(model_gg_covariates, model_gg_covariates_v2,
  type = "text",    # Use "html" for HTML output or "latex" for LaTeX
  title = "Regression Results for Good and Gather Score",
  covariate.labels = c("Treatment", "log(Age)", "Gender", "Chameleon Awareness", "Good and Gather Awareness"),
  star.cutoffs = c(0.10, 0.05, 0.01, 0.001),    # Significance stars
  out = "regression_table.txt") # Optional: Save output to a text file
```

```
##
## Regression Results for Good and Gather Score
## =====
##                                     Dependent variable:
##                                     -----
##                                     good_and_gather_score
##                                     (1)          (2)
## -----
## Treatment                        -0.568*      -0.513*
##                                (0.287)      (0.294)
##
## log(Age)                        1.303***      1.130**
##                                (0.428)      (0.442)
##
## Gender                          0.385         0.254
##                                (0.276)      (0.289)
##
## Chameleon Awareness             -1.887**      -2.097**
##                                (0.945)      (0.960)
##
## Good and Gather Awareness        0.932*        0.951*
##                                (0.475)      (0.493)
##
## chameleon_awarenessYes, Positive -0.617        -0.526
##                                (0.401)      (0.421)
##
## good_and_gather_awarenessYes, Negative -0.856
##                                (1.350)
##
## good_and_gather_awarenessYes, Neutral -0.543
```

```

##                                     (0.350)
##
## good_and_gather_awarenessYes, Positive      -0.147
##                                             (0.391)
##
## Constant          -0.524          0.314
##                   (1.512)        (1.605)
## -----
## Observations          90          90
## R2                    0.229        0.254
## Adjusted R2           0.174        0.170
## Residual Std. Error    1.301 (df = 83)    1.304 (df = 80)
## F Statistic           4.116*** (df = 6; 83) 3.032*** (df = 9; 80)
## =====
## Note:                                     *p<0.1; **p<0.05; ***p<0.01

```