UTILITY's Home Energy Report Analysis

Pedro Huet López

2024-04-07

Instructions:

There is a survey instrument and a file of survey results. Please use the data to address the following questions:

1. What percentage of all treatment group (print and email) customers responded that they read their reports in detail?

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
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
```

```
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
library(readxl)

X2023_Quant_Assignment_Data <- read_excel("C:/Users/pedro/Downloads/2023 Quant Assignment Data.x
lsx")
View(X2023_Quant_Assignment_Data)

#Visualize the full group of interest

treatment_counts <- table(X2023_Quant_Assignment_Data$TREATMENT)
total_count <- treatment_counts["Print"] + treatment_counts["Email"]

print(treatment_counts["Print"])
```

```
## Print
##   156
```

```
print(treatment_counts["Email"])
```

```
## Email
##   148
```

```
print(total_count)
```

```
## Print
##   304
```

The data shows that the total treatment group (print and email) is comprised of 304 customers. In order to calculate the size of the subgroup of treated customers who read the indications in detail, we are only missing the total number number of treated customers who read the text in detail.

```
# Count Variable of interest: "Reading Print in detail Print" G1
count_G1 <- sum(X2023_Quant_Assignment_Data$G1 == 1, na.rm = TRUE)

print(paste("Read Print in detail:", count_G1))
```

```
## [1] "Read Print in detail: 56"
```

```
# Count Variable of interest: "Reading Email in detail Email" H1
count_H1 <- sum(X2023_Quant_Assignment_Data$H1 == 1, na.rm = TRUE)

print(paste("Read Email in detail:", count_H1))
```

```
## [1] "Read Email in detail: 36"
```

The data shows that 56 customers who received the print report read it in detail, while 36 customers who received the email read it in detail.

*Note: due to social desirability bias, it is possible that some users are exaggerating the detail with which they read the reports, so its reasonable to assume these figures are optimistic.

```
#Calculations of percentages.
percentage <- ((count_G1 + count_H1) / total_count) * 100

percentage_rounded <- round(percentage, 2)
percentage_with_sign <- paste(percentage_rounded, "%", sep = "")

print(paste("The percentage of all treatment group (print and email) customers that responded th
ey read their reports in detail was", percentage_with_sign))
```

```
## [1] "The percentage of all treatment group (print and email) customers that responded they re
ad their reports in detail was 30.26%"
```

2. Describe any differences among the three groups of customers in how they understand their electric bill and usage (E4, E6, E7)?

```
print_distribution <- function(data, column_name) {
  distribution <- table(data[[column_name]])
  cat("Distribution of values in", column_name, ":\n")
  print(distribution)
}

columns_to_print <- c("E4", "E6", "E7")

invisible(lapply(columns_to_print, function(col) print_distribution(X2023_Quant_Assignment_Data,
col)))
```

```
## Distribution of values in E4 :
##
##   1   2
## 351  76
## Distribution of values in E6 :
##
##   1   2   3   4   5   6   7  98
##  10  30 301  48  10   9  16   2
## Distribution of values in E7 :
##
##   1   2   3   4   5
##  56 135 142  68  25
```

Before proceeding to disaggregate the data, it would make sense to process it to make easily quantifiable estimates. Specifically, for the question regarding frequency with which customers monitor their electricity use or electricity bill expenses (E6), non respondants were dropped from estimations because the variable was coded as 98, which would affect the variable's distribution. Additionally, since there were only 2 observations that were in this situation, dropping them shouldn't modify the rest of the results in evene the slightest manner.

```r
#We drop void answers.
X2023_Quant_Assignment_Data <- X2023_Quant_Assignment_Data[X2023_Quant_Assignment_Data$E6 != 98,
]

X2023_Quant_Assignment_Data$E4[X2023_Quant_Assignment_Data$E4 == 2] <- 0

invisible(lapply(columns_to_print, function(col) print_distribution(X2023_Quant_Assignment_Data,
col)))
```

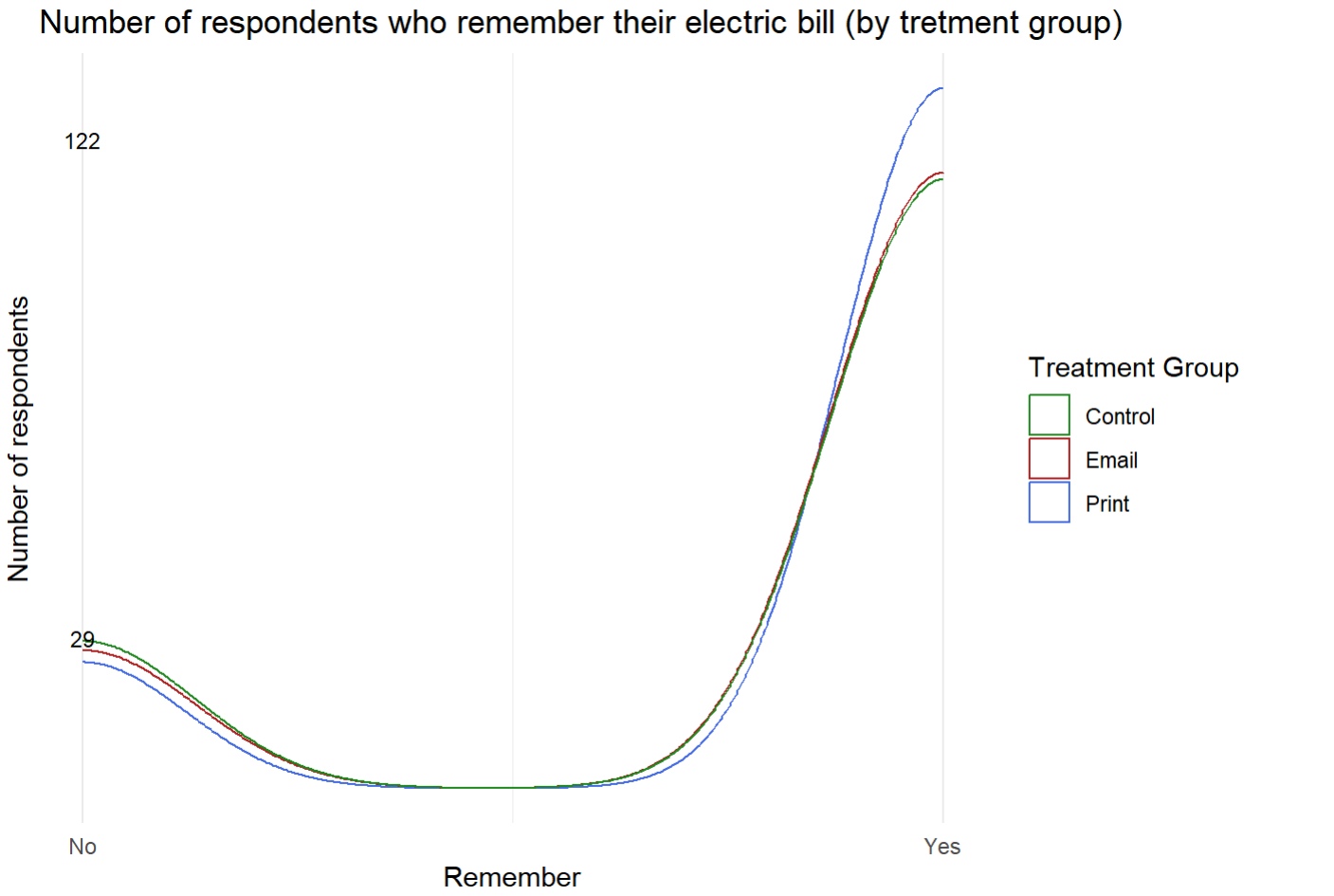
```
## Distribution of values in E4 :
##
##   0   1
##  75 349
## Distribution of values in E6 :
##
##   1   2   3   4   5   6   7
##  10  30 301  48  10   9  16
## Distribution of values in E7 :
##
##   1   2   3   4   5
##  56 134 141  68  25
```

```r
# Counts of variables by treatment group

summarize_variable <- function(variable, data) {
  table_var <- table(data$TREATMENT, data[[variable]])
  prop_table <- prop.table(table_var, margin = 1)
  prop_table_rounded <- round(prop_table, digits = 2)  # Rounding up to 2 decimals
  return(list(counts = table_var, percentages = prop_table_rounded))
}

variables <- c("E4", "E6", "E7")
tables <- lapply(variables, summarize_variable, data = X2023_Quant_Assignment_Data)

for (i in seq_along(variables)) {
  cat("Counts and Percentages of", variables[i], "by Treatment Group:\n")
  cat("Counts:\n")
  print(tables[[i]]$counts)
  cat("Percentages:\n")
  print(tables[[i]]$percentages)
}
```

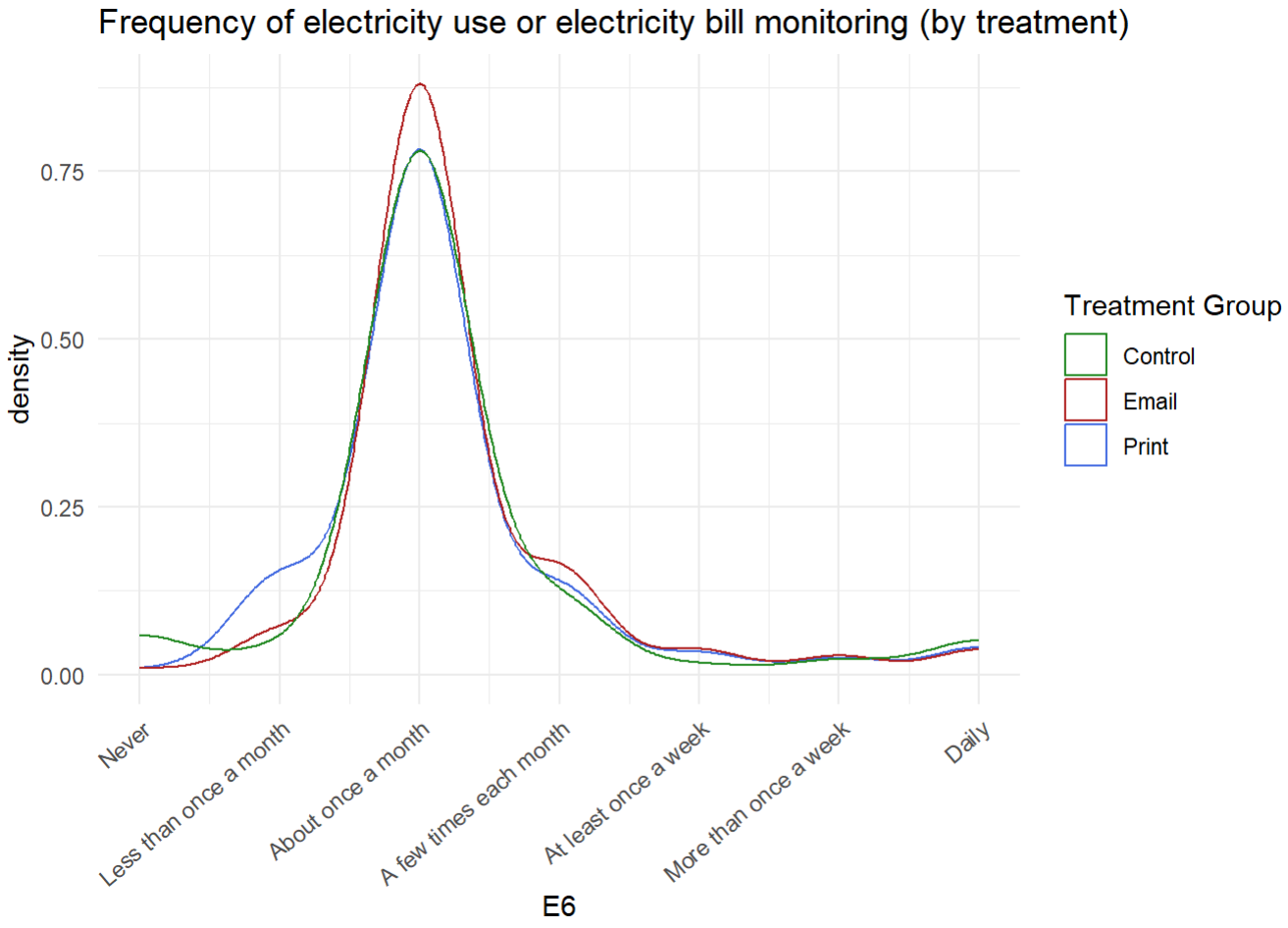
```
## Counts and Percentages of E4 by Treatment Group:
## Counts:
##
##             0   1
##   Control  29 120
##   Email    24 107
##   Print    22 122
## Percentages:
##
##              0    1
##   Control 0.19 0.81
##   Email   0.18 0.82
##   Print   0.15 0.85
## Counts and Percentages of E6 by Treatment Group:
## Counts:
##
##             1   2   3   4   5   6   7
##   Control   8   5 109  15   2   3   7
##   Email     1   7  95  17   4   3   4
##   Print     1  18  97  16   4   3   5
## Percentages:
##
##              1    2    3    4    5    6    7
##   Control 0.05 0.03 0.73 0.10 0.01 0.02 0.05
##   Email   0.01 0.05 0.73 0.13 0.03 0.02 0.03
##   Print   0.01 0.12 0.67 0.11 0.03 0.02 0.03
## Counts and Percentages of E7 by Treatment Group:
## Counts:
##
##            1  2  3  4  5
##   Control 27 55 46 13  8
##   Email   15 30 48 29  9
##   Print   14 49 47 26  8
## Percentages:
##
##              1    2    3    4    5
##   Control 0.18 0.37 0.31 0.09 0.05
##   Email   0.11 0.23 0.37 0.22 0.07
##   Print   0.10 0.34 0.33 0.18 0.06
```

The Plots look as follows:

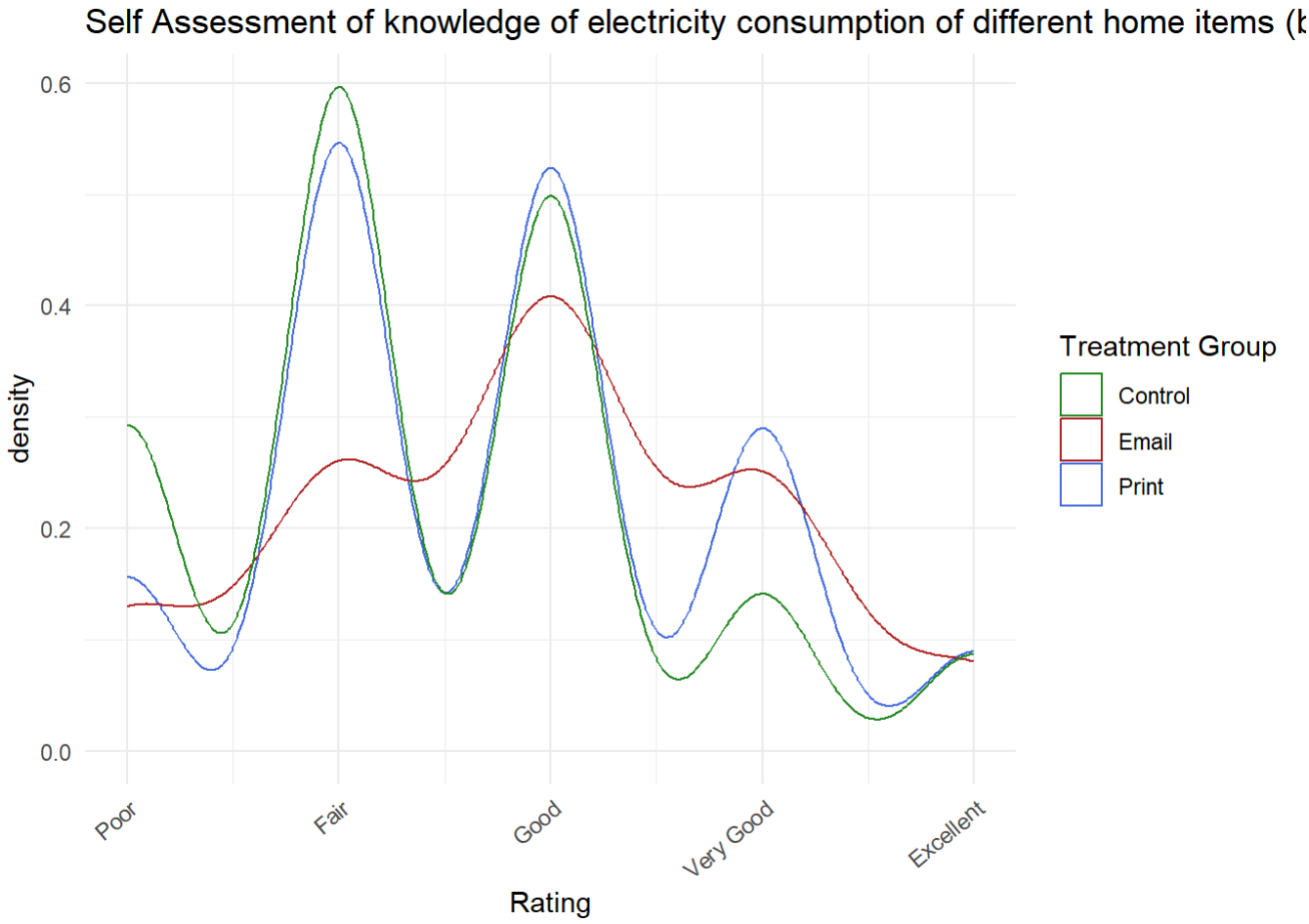
```
print_data <- na.omit(X2023_Quant_Assignment_Data[X2023_Quant_Assignment_Data$TREATMENT == "Prin
t", c("E4", "E6", "E7")])
email_data <- na.omit(X2023_Quant_Assignment_Data[X2023_Quant_Assignment_Data$TREATMENT == "Emai
l", c("E4", "E6", "E7")])
control_data <- na.omit(X2023_Quant_Assignment_Data[X2023_Quant_Assignment_Data$TREATMENT == "Co
ntrol", c("E4", "E6", "E7")])

# Plot for variable E4
ggplot() +
  geom_density(aes(x = E4, color = "Print"), data = print_data) +
  geom_density(aes(x = E4, color = "Email"), data = email_data) +
  geom_density(aes(x = E4, color = "Control"), data = control_data) +
  labs(title = "Number of respondents who remember their electric bill (by tretment group)", x =
"Remember", y = "Number of respondents", color = "Treatment Group") +
  scale_color_manual(values = c("Print" = "royalblue", "Email" = "firebrick", "Control" = "fores
tgreen")) +
  scale_x_continuous(breaks = c(0, 1), labels = c("No", "Yes")) +
  annotate("text", x = 0, y = 0.5, label = 29, vjust = -1, color = "black", size = 3) +
  annotate("text", x = 0, y = 2.5, label = 122, vjust = -1, color = "black", size = 3) +
  scale_y_continuous(breaks = NULL) +
  theme_minimal()
```



```
# Plot for variable E6
ggplot() +
  geom_density(aes(x = E6, color = "Print"), data = print_data) +
  geom_density(aes(x = E6, color = "Email"), data = email_data) +
  geom_density(aes(x = E6, color = "Control"), data = control_data) +
  labs(title = "Frequency of electricity use or electricity bill monitoring (by treatment)", x =
"E6", color = "Treatment Group") +
  scale_color_manual(values = c("Print" = "royalblue", "Email" = "firebrick", "Control" = "fores
tgreen")) +
  scale_x_continuous(breaks = 1:7, labels = c("Never",
                                              "Less than once a month",
                                              "About once a month",
                                              "A few times each month",
                                              "At least once a week",
                                              "More than once a week",
                                              "Daily")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 40, hjust = 1))
```



```
# Plot for variable E7
ggplot() +
  geom_density(aes(x = E7, color = "Print"), data = print_data) +
  geom_density(aes(x = E7, color = "Email"), data = email_data) +
  geom_density(aes(x = E7, color = "Control"), data = control_data) +
  labs(title = "Self Assessment of knowledge of electricity consumption of different home items
(by treatment)", x = "Rating", color = "Treatment Group") +
  scale_color_manual(values = c("Print" = "royalblue", "Email" = "firebrick", "Control" = "fores
tgreen")) +
    scale_x_continuous(breaks = 1:5, labels = c("Poor",
                                                "Fair",
                                                "Good",
                                                "Very Good",
                                                "Excellent")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 40, hjust = 1))
```



Regression analysis results show the following trends:

```
dummy_data <- dummyVars("~ TREATMENT", data = X2023_Quant_Assignment_Data)
dummy_variables <- predict(dummy_data, newdata = X2023_Quant_Assignment_Data)

X2023_Quant_Assignment_Datn_Ordinal <- cbind(X2023_Quant_Assignment_Data, dummy_variables)
```

```
###From a Print Approach##

model <- lm(E4 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordinal)

logit_model <- glm(E4 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Or
dinal, family = binomial(link = "logit"))

summary(logit_model)
```

```
##
## Call:
## glm(formula = E4 ~ TREATMENTControl + TREATMENTEmail, family = binomial(link = "logit"),
##     data = X2023_Quant_Assignment_Datn_Ordinal)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.7130     0.2316   7.395 1.41e-13 ***
## TREATMENTControl   -0.2928     0.3106  -0.943    0.346
## TREATMENTEmail     -0.2182     0.3235  -0.674    0.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 423  degrees of freedom
## Residual deviance: 394.76  on 421  degrees of freedom
##   (30 observations deleted due to missingness)
## AIC: 400.76
##
## Number of Fisher Scoring iterations: 4
```

```
summary(model)
```

```
##
## Call:
## lm(formula = E4 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8472  0.1528  0.1832  0.1946  0.1946
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.84722    0.03188  26.579   <2e-16 ***
## TREATMENTControl -0.04185    0.04470  -0.936     0.35
## TREATMENTEmail   -0.03043    0.04618  -0.659     0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3825 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.0022, Adjusted R-squared:  -0.00254
## F-statistic: 0.4642 on 2 and 421 DF,  p-value: 0.6289
```

```
model2 <- lm(E6 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordinal)

summary(model2)
```

```
##
## Call:
## lm(formula = E6 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3130 -0.3130 -0.2349 -0.2292  3.7708
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.229167   0.087710  36.816   <2e-16 ***
## TREATMENTControl 0.005733   0.122996   0.047    0.963
## TREATMENTEmail   0.083810   0.127081   0.660    0.510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.001274,   Adjusted R-squared:  -0.003471
## F-statistic: 0.2685 on 2 and 421 DF,  p-value: 0.7647
```

```
model3 <- lm(E7 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordinal)

summary(model3)
```

```
##
## Call:
## lm(formula = E7 ~ TREATMENTControl + TREATMENTEmail, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90076 -0.75694  0.09924  0.53691  2.53691
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.75694    0.08835  31.204   <2e-16 ***
## TREATMENTControl -0.29386    0.12390  -2.372   0.0182 *
## TREATMENTEmail    0.14382    0.12801   1.123   0.2619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.02895,    Adjusted R-squared:  0.02434
## F-statistic: 6.275 on 2 and 421 DF,  p-value: 0.002063
```

```
#From an Email Approach

model1m <- lm(E4 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordina
l)

logit_modelm <- glm(E4 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_O
rdinal, family = binomial(link = "logit"))

summary(model1m)
```

```
##
## Call:
## lm(formula = E4 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8472  0.1528  0.1832  0.1946  0.1946
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.81679    0.03342  24.440   <2e-16 ***
## TREATMENTControl -0.01142    0.04581  -0.249    0.803
## TREATMENTPrint    0.03043    0.04618   0.659    0.510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3825 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.0022, Adjusted R-squared:  -0.00254
## F-statistic: 0.4642 on 2 and 421 DF,  p-value: 0.6289
```

```
summary(logit_modelm)
```

```
##
## Call:
## glm(formula = E4 ~ TREATMENTControl + TREATMENTPrint, family = binomial(link = "logit"),
##     data = X2023_Quant_Assignment_Datn_Ordinal)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.49478    0.22586   6.618 3.64e-11 ***
## TREATMENTControl -0.07458    0.30631  -0.243    0.808
## TREATMENTPrint    0.21820    0.32352   0.674    0.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 423  degrees of freedom
## Residual deviance: 394.76  on 421  degrees of freedom
##   (30 observations deleted due to missingness)
## AIC: 400.76
##
## Number of Fisher Scoring iterations: 4
```

```
model2m <- lm(E6 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordina
l)

summary(model2m)
```

```
##
## Call:
## lm(formula = E6 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3130 -0.3130 -0.2349 -0.2292  3.7708
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.31298    0.09196  36.027   <2e-16 ***
## TREATMENTControl -0.07808    0.12606  -0.619    0.536
## TREATMENTPrint   -0.08381    0.12708  -0.660    0.510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.001274,   Adjusted R-squared:  -0.003471
## F-statistic: 0.2685 on 2 and 421 DF,  p-value: 0.7647
```

```
model3m <- lm(E7 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordina
l)

summary(model3m)
```

```
##
## Call:
## lm(formula = E7 ~ TREATMENTControl + TREATMENTPrint, data = X2023_Quant_Assignment_Datn_Ordin
al)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90076 -0.75694  0.09924  0.53691  2.53691
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.90076    0.09263  31.314  < 2e-16 ***
## TREATMENTControl -0.43768    0.12699  -3.447 0.000625 ***
## TREATMENTPrint   -0.14382    0.12801  -1.123 0.261877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 421 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.02895,    Adjusted R-squared:  0.02434
## F-statistic: 6.275 on 2 and 421 DF,  p-value: 0.002063
```

Descriptive Statistics and linear regression results seem to imply that the main difference among the 3 treatment groups mostly lies with respondents' perceived knowledge of electricity consumption of their different home items.

1.According to the results of the plots and percentages, there is only a slight difference in remembering electrical bill expenses between costumers who don't receive email or print (81% remember)information compared to receiving email (82% remember) and print (85% remember). When performing simple regression analysis, it would seem customers who receive print more closely remembering their expenses, but these differences are miniscule (between 3% to 4% more retention) and the coefficients aren't statistically significant.

2.With regards to the frequency with which customers monitor their electricity use or electricity bill expenses, the three groups also seemed to behave similarly: there is no major differences between the monitoring frequency of customers who don't receive email or print information, as they mostly check it once a month (between 67% to 73% per group) or more than once a month (between 11% to 13%). Simple regression analysis shows that, among these three customer groups, those that receive information via email, on average, tend to be the ones that monitor their electricity use more. However, this difference is also small (between 7% to 8% more awareness) and the coefficients aren't statistically significant.

3.- Finally, the self assessment of electricity consumption of customers who receive email or print information seem to rank higher than those who don't. The plots and regressions seem to show that customers who don't receive information, on average, self assess between 0.29 (Print) and 0.43 (Email) points less (in a scale of 1 to 5) than customers who receive the information. These differences are statistically significant at significance levels above, at least, 5%. Comparing the differences between the email and print groups, the email group scored slightly higher (0.14 points more), but this difference wasn't statistically significant.

Overall, these differences seem to imply that information being sent to consumers via email and print makes a noticeable difference compared to not receiving it in terms of their self assessment of electricity consumption by household item. Additionally, since the evidence shows that the effect of emailing information was at least as good as giving print information, it would make sense to choose emailing information instead of sending it print if this meant less expenses (administrative costs and material) and a more sustainable and eco-friendly approach.

3. Discuss if/how overall satisfaction with UTILITY varies by income level?

Based on this question, the best satisfaction measure of UTILITY'S services seems to be question E1: "Taking into consideration all aspects of your utility service experience, how would you rate UTILITY overall? Please use a 1 to 10 scale where 1 is 'poor' and 10 is 'excellent'."

With regards to income, the best measure among the survey questions seems to be the question J4: "Which of the following categories best represents your total annual household income before taxes? Please tell me when I get to your range." Although this question seems to measure income directly, it is important to note that there is a chance it could introduce bias in the response (i.e. some respondent's might feel nervous about sharing this information and might offer smaller figures), so the results should be taken with some skepticism.

In any case, a measure of this could be calculated via a simple linear regression model. To do so, however, it is necessary to synchronize the data that comes from different sources (different sheets in an excel file), this will be done by mapping them to the User Id's they hhave in common (commonly known as a "crosswalk").

```
#Crosswalk to perform regression analysis:

first_sheet <- read_excel("C:/Users/pedro/Downloads/2023 Quant Assignment Data.xlsx", sheet = 1)
second_sheet <- read_excel("C:/Users/pedro/Downloads/2023 Quant Assignment Data.xlsx", sheet =
2)

combined_data <- merge(first_sheet, second_sheet, by = "RespondentID", all = TRUE)

modified_data <- combined_data %>%
  mutate(J4 = ifelse(J4 > 10, NA, J4))
```

```
satis_model <- lm(E1 ~ J4, data = modified_data)
summary(satis_model)
```

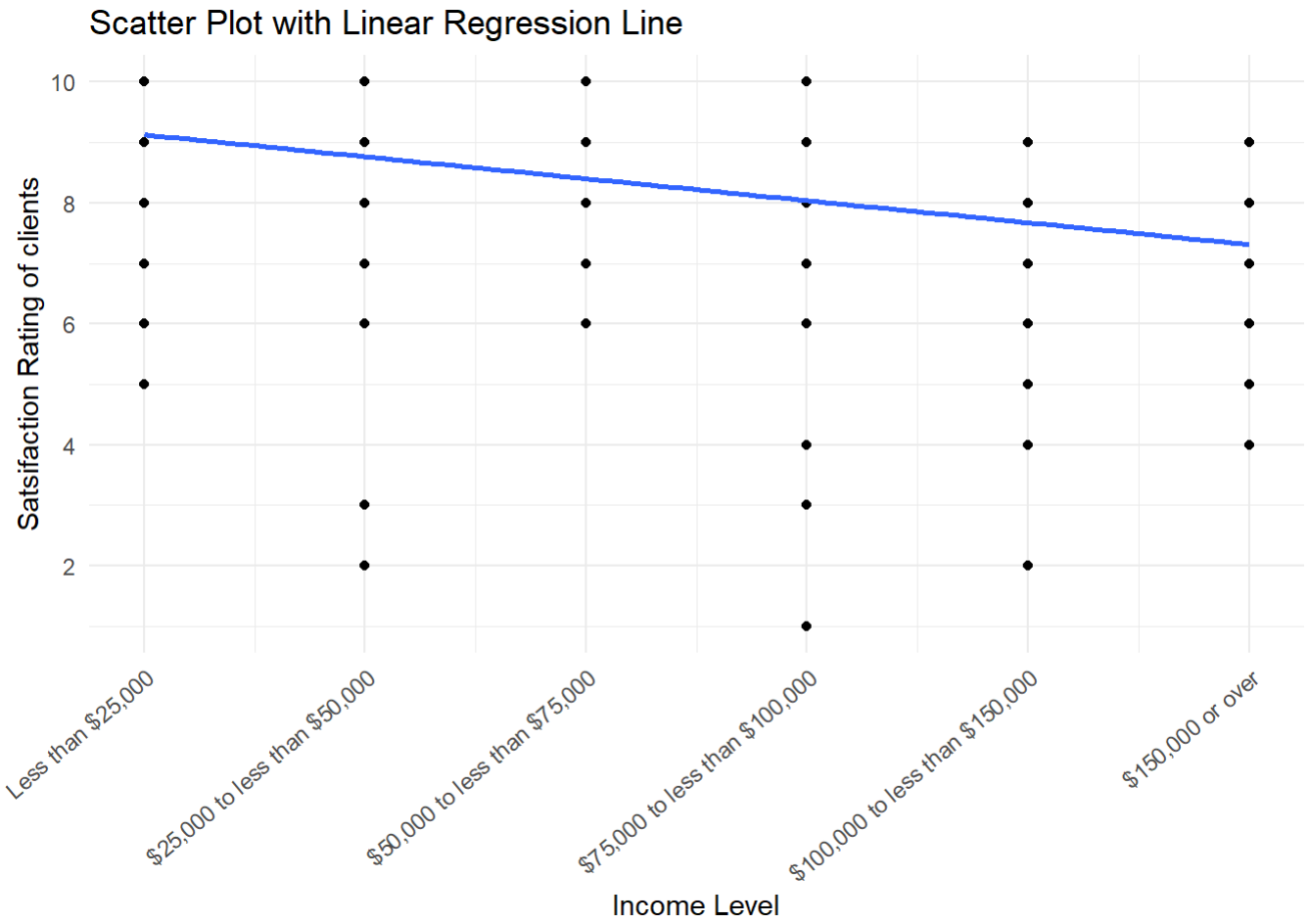
```
##
## Call:
## lm(formula = E1 ~ J4, data = modified_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0399 -0.6763  0.3237  1.2329  1.9601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.49433    0.20363  46.626  < 2e-16 ***
## J4          -0.36361    0.05096  -7.135 5.63e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.471 on 349 degrees of freedom
##   (112 observations deleted due to missingness)
## Multiple R-squared:  0.1273, Adjusted R-squared:  0.1248
## F-statistic: 50.91 on 1 and 349 DF,  p-value: 5.634e-12
```

```
custom_labels <- c("Less than $25,000",
                   "$25,000 to less than $50,000",
                   "$50,000 to less than $75,000",
                   "$75,000 to less than $100,000",
                   "$100,000 to less than $150,000",
                   "$150,000 or over")

ggplot(modified_data, aes(x = J4, y = E1)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Income Level", y = "Satsifaction Rating of clients") +
  ggtitle("Scatter Plot with Linear Regression Line") +
  scale_x_continuous(breaks = 1:6, labels = custom_labels) +
  scale_y_continuous(breaks = pretty(modified_data$E1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 40, hjust = 1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 112 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 112 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
mean_E1 <- mean(modified_data$E1, na.rm = TRUE)
print(mean_E1)
```

```
## [1] 7.983759
```

According to these results, it would seem that income level could have an effect on a customer's reported satisfaction with UTILITY. In this case, the higher customers' income, the less they tend to be satisfied with UTILITY Services. Specifically, an increase of income level (between 25,000 USD to 50,000 USD) results in a 0.34 unit (in a scale of 1 to 10) decrease in customer satisfaction. This result is statistically significant to a significance level of 0.001%, which means that the evidence at hand makes it too unlikely to think that this pattern is a product of mere chance. The R-squared value of 0.12 indicates that the income level of clients, by itself, seems to account for at least 12% of the fluctuations we see on customer satisfaction, which is quite high for a single variable. All in all, the evidence suggests that the income level of customers could have a noticeable moderate sized negative effect on customer satisfaction with UTILITY.

Of course, this result should be taken in context:

Customer satisfaction with UTILITY was, in general, very already high (on average, customers rated it around 8.0 out of 10.0), so this result does not mean a structural or grave issue the company is facing. A possible interpretation is that this pattern is a product of expectations: as customers have higher incomes, they tend to expect more out of services they receive, as they are likely able to afford and choose better services during their daily lives. Therefore, when these customers with higher expectations receive UTILITY's services, they could feel some disappointment that they perceive them to not be on par with other premium services of the same or other industries they have already recieved.

A final implication is that this could be a business or market opportunity for UTILITY: perhaps UTILITY could offer services that are more thorough or of a higher quality, so that customers of a higher income level could self select into these and report a higher satisfaction level. This could lead to high profits for the organization and higher satisfaction from clients.