Lab 1, Short Questions

Contents


```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
library(patchwork)
```

1 Strategic Placement of Products in Grocery Stores (5 points)

These questions are taken from Question 12 of chapter 3 of the textbook(Bilder and Loughin's "Analysis of Categorical Data with R.

> *In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item—breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular boxes of cereals on specific shelves, grocery stores may better attract customers to them. To investigate this further, a random sample of size 10 was taken from each of four shelves at a Dillons grocery store in Manhattan, KS. These data are given in the* cereal_dillons.csv *file. The response variable is the shelf number, which is numbered from bottom (1) to top (4), and the explanatory variables are the sugar, fat, and sodium content of the cereals.*

```
setwd("~/source_code/W271_Lab1")
cereal <- read_csv("./data/short-questions/cereal_dillons.csv")
```

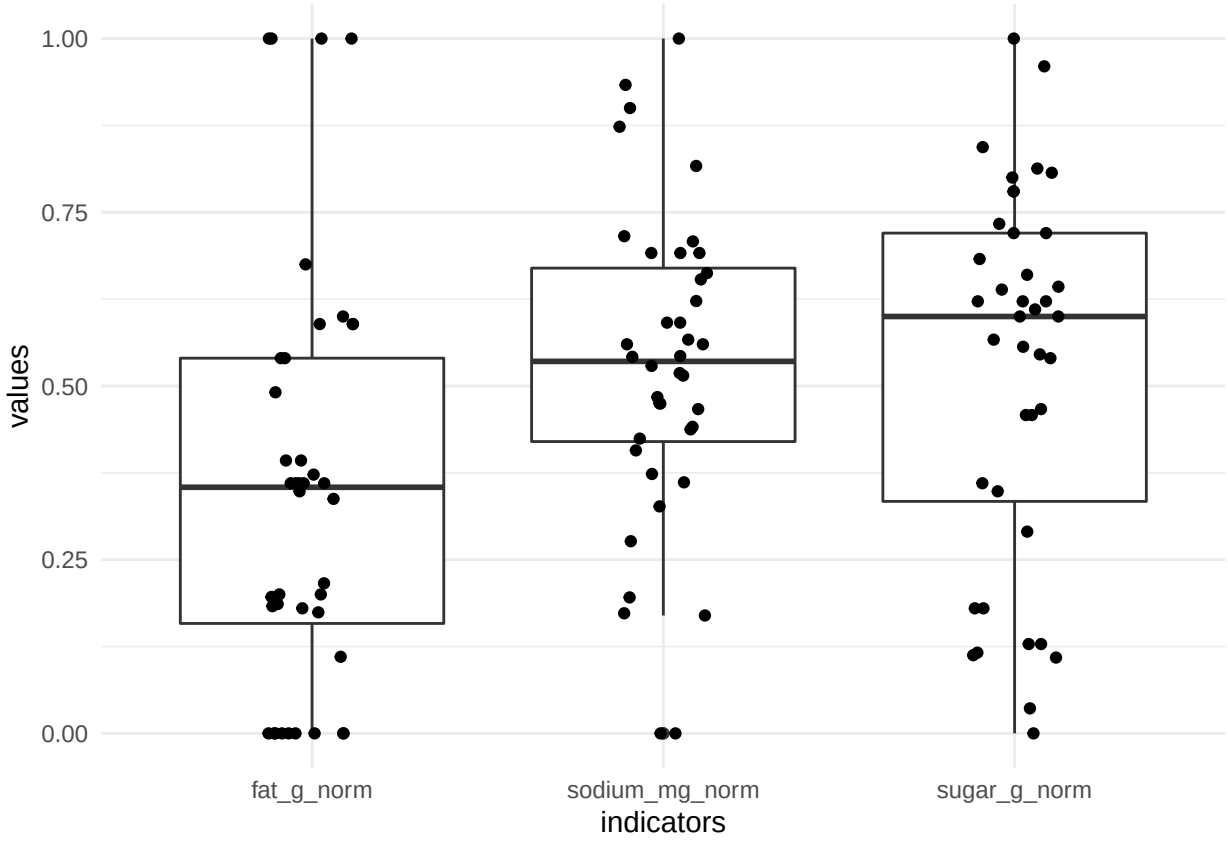
1.1 Recode Data

(1 point) The explanatory variables need to be reformatted before proceeding further (sample code is provided in the textbook). First, divide each explanatory variable by its serving size to account

for the different serving sizes among the cereals. Second, rescale each variable to be within 0 and 1. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables. Also, construct a parallel coordinates plot for the explanatory variables and the shelf number. Discuss whether possible content differences exist among the shelves.

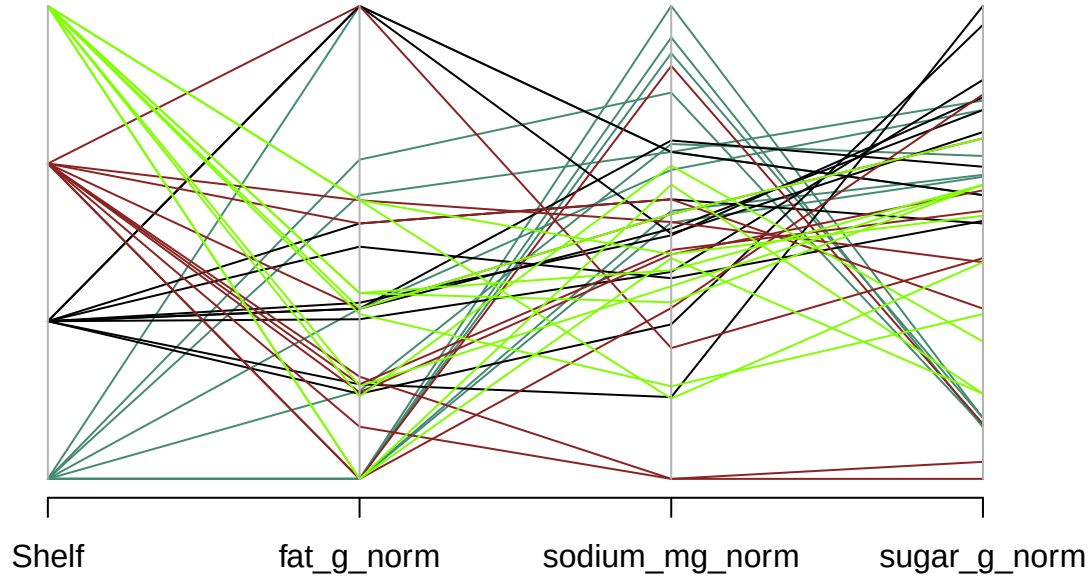
```
col_trans_list <- c("fat_g", "sugar_g", "sodium_mg")


for (col in col_trans_list) {
  col_per <- cereal[col] / cereal["size_g"]
  cereal[paste(col, "_norm", sep = "")] <- (col_per - min(col_per)) / (max(col_per) - min(col_
}
```

```
cereal %>%
  pivot_longer(cols = 8:10, names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = indicators, y = values)) +
  geom_boxplot() +
  geom_jitter(width = 0.125) +
  theme_minimal()
```



'Fill in: What do you observe in these boxplots?'

```
MASS::parcoord(cereal[c("Shelf", "fat_g_norm", "sodium_mg_norm", "sugar_g_norm")], col = colors
```



'Fill in: What do you observe in these parallel coordinates plots?'

Fill in: Do content differences exist between the shelves?'

1.2 Evaluate Ordinal vs. Categorical

(1 point) The response has values of $1, 2, 3$, and 4. Explain under what setting would it be desirable to take into account ordinality, and whether you think that this setting occurs here. Then estimate a suitable multinomial regression model with linear forms of the sugar, fat, and sodium variables. Perform LRTs to examine the importance of each explanatory variable. Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables).

Fill in: What do you think about ordinal data?'

```
library(nnet)
model_cereal_shelves_linear <- multinom(
  formula = Shelf ~ fat_g_norm +
    sugar_g_norm +
    sodium_mg_norm,
  data = cereal
)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 37.329384
## iter  20 value 33.775257
## iter  30 value 33.608495
## iter  40 value 33.596631
## iter  50 value 33.595909
## iter  60 value 33.595564
## iter  70 value 33.595277
## iter  80 value 33.595147
## final  value 33.595139
## converged
```

```
model_cereal_shelves_quadratic <- multinom(
  formula = Shelf ~ fat_g_norm +
    sugar_g_norm +
    sodium_mg_norm +
    fat_g_norm:sugar_g_norm +
    fat_g_norm:sodium_mg_norm +
    sodium_mg:sugar_g_norm +
    sodium_mg:sugar_g_norm:fat_g_norm,
  data = cereal
)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 38.867436
## iter  20 value 30.619057
## iter  30 value 27.228569
## iter  40 value 26.252054
## iter  50 value 25.788200
## iter  60 value 25.733143
## iter  70 value 25.523018
## iter  80 value 25.347699
## iter  90 value 25.268623
## iter 100 value 25.189962
## final  value 25.189962
## stopped after 100 iterations
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.2
```

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

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

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

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
lrt_cereal_main_effects <- car::Anova(model_cereal_shelves_linear)
lrt_cereal_main_effects
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                LR Chisq Df Pr(>Chisq)
## fat_g_norm       5.2836  3     0.1522
## sugar_g_norm    22.7648  3  4.521e-05 ***
## sodium_mg_norm  26.6197  3  7.073e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrt_cereal_quadratic_effects <- car::Anova(model_cereal_shelves_quadratic)
lrt_cereal_quadratic_effects
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                                    LR Chisq Df Pr(>Chisq)
## fat_g_norm                            5.524  3  0.1371874
## sugar_g_norm                         19.252  3  0.0002424 ***
## sodium_mg_norm                       35.008  3  1.213e-07 ***
## fat_g_norm:sugar_g_norm               3.536  3  0.3160843
## fat_g_norm:sodium_mg_norm             3.738  3  0.2912061
## sugar_g_norm:sodium_mg                5.243  3  0.1548738
## fat_g_norm:sugar_g_norm:sodium_mg     3.529  3  0.3169753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

'Fill in: Write about what you learn as a result of these tests, using inline code evaluation.'

1.3 Where do you think Apple Jacks will be placed?

(1 point) Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed toward children. For a serving size of 28 grams, its sugar content is 12 grams, fat content is 0.5 grams, and sodium content is 130 milligrams. Estimate the shelf probabilities for Apple Jacks.

```
# Estimate new model that removes non-significant fat variable
model_cereal_shelves_trim <- multinom(formula = Shelf ~ sugar_g_norm + sodium_mg_norm, data = c
```

```
## # weights:  16 (9 variable)
## initial  value 55.451774
## iter  10 value 39.834294
## iter  20 value 36.269828
## iter  30 value 36.248421
## iter  40 value 36.241637
```

```
## iter  50 value 36.238788
## iter  60 value 36.237394
## iter  70 value 36.237065
## iter  80 value 36.236920
## iter  80 value 36.236919
## iter  80 value 36.236919
## final  value 36.236919
## converged
```

```r
# Create a datframe with Apple Jack data normalized to the same scale as the training data set
app_jack <- data.frame(size_g = 28, sugar_g = 12, fat_g = 0.5, sodium_mg = 130)
for (col in col_trans_list) {
  col_per <- app_jack[col] / app_jack["size_g"]
  ref_col <- cereal[col] / cereal["size_g"]
  app_jack[paste(col, "_norm", sep = "")] <- (col_per - min(ref_col)) / max(ref_col)
}

# Estimate placement of Apple Jack
aj_shelf_probs_trim <- predict(model_cereal_shelves_trim, newdata = app_jack, type = "probs")
shelf_trim <- aj_shelf_probs_trim[which.max(aj_shelf_probs_trim)]
shelf_trim
```

```
##        2
## 0.590739
```

```r
aj_shelf_probs <- predict(model_cereal_shelves_linear, newdata = app_jack, type = "probs")
shelf <- aj_shelf_probs[which.max(aj_shelf_probs)]
shelf
```

```
##         2
## 0.4719426
```

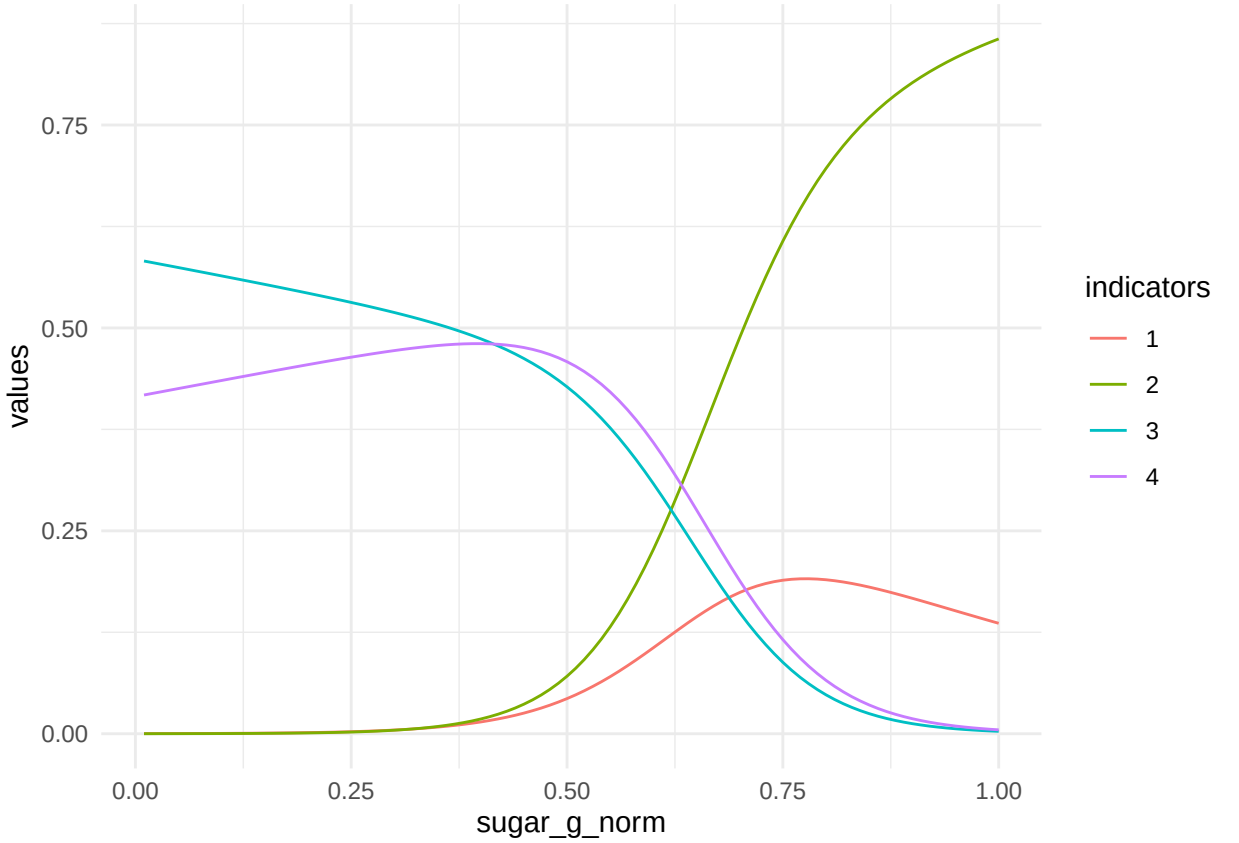
'Fill this in: Where does your model predict apple jacks will be placed?'

1.4 Figure 3.3

(1 point) Construct a plot similar to Figure 3.3 where the estimated probability for a shelf is on the *y-axis* and the sugar content is on the *x-axis*. Use the mean overall fat and sodium content as the corresponding variable values in the model. Interpret the plot with respect to sugar content.

```r
fat_mean <- mean(cereal$fat_g_norm)
na_mean <- mean(cereal$sodium_mg_norm)

df_to_plot <- data.frame(fat_g_norm = rep(fat_mean, times = 100), sodium_mg_norm = rep(na_mean
df_to_plot <- cbind(df_to_plot, predict(model_cereal_shelves_linear, newdata = df_to_plot, type

shelf_vs_sugar_plot <- df_to_plot %>%
  pivot_longer(cols = c(4:7), names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = sugar_g_norm, y = values, colour = indicators)) +
  geom_line() +
  theme_minimal()
```

```
shelf_vs_sugar_plot
```



'Fill this in: What message does your plot give?'

1.5 Odds ratios

(1 point) Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretations back to the plots constructed for this exercise.

```
odds_ratios <- "fill this in"
```

'Fill this in: What do you learn about each of these variables?'

2 Alcohol, self-esteem and negative relationship interactions (5 points)

Read the example **'Alcohol Consumption'** in chapter 4.2.2 of the textbook(Bilder and Loughin's "Analysis of Categorical Data with R). This is based on a study in which moderate-to-heavy drinkers (defined as at least 12 alcoholic drinks/week for women, 15 for men) were recruited to keep a daily record of each drink that they consumed over a 30-day study period. Participants also completed a variety of rating scales covering daily events in their lives and items related to self-esteem. The data are given in the *DeHartSimplified.csv* data set. Questions 24-26 of chapter 3 of the textbook also relate to this data set and give definitions of its variables: the number of drinks consumed

(`numall`), positive romantic-relationship events (`prel`), negative romantic-relationship events (`nrel`), age (`age`), trait (long-term) self-esteem (`rosn`), state (short-term) self-esteem (`state`).

The researchers stated the following hypothesis:

> *We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem.*

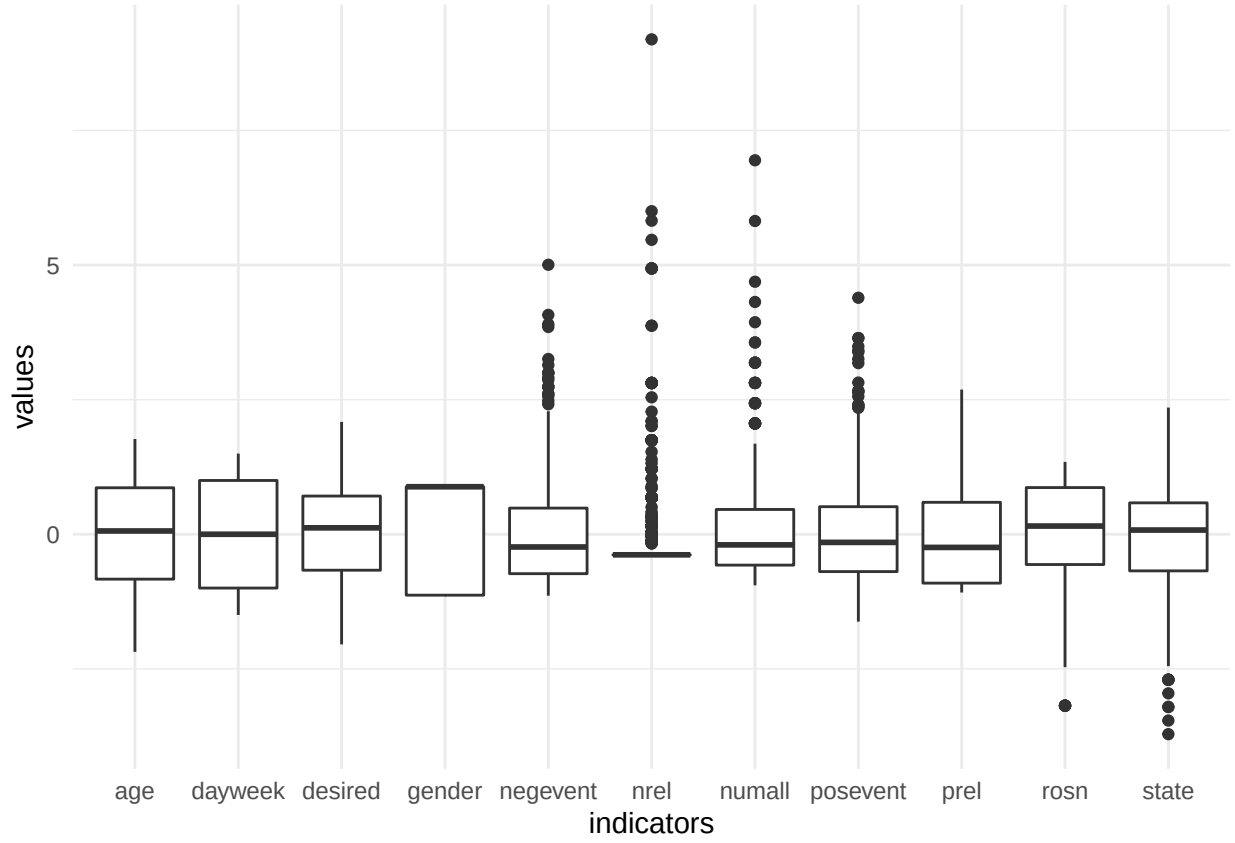
```
drinks <- read_csv("../data/short-questions/DeHartSimplified.csv")
```

2.1 EDA

(2 points) Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. Address the reasons for limiting the study to observations from only one day.

```
drinks_scaled <- as.data.frame(scale(drinks))
drinks_scaled%>%
  pivot_longer(cols = 3:13, names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = indicators, y = values)) +
  geom_boxplot() +
  theme_minimal()
```

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```



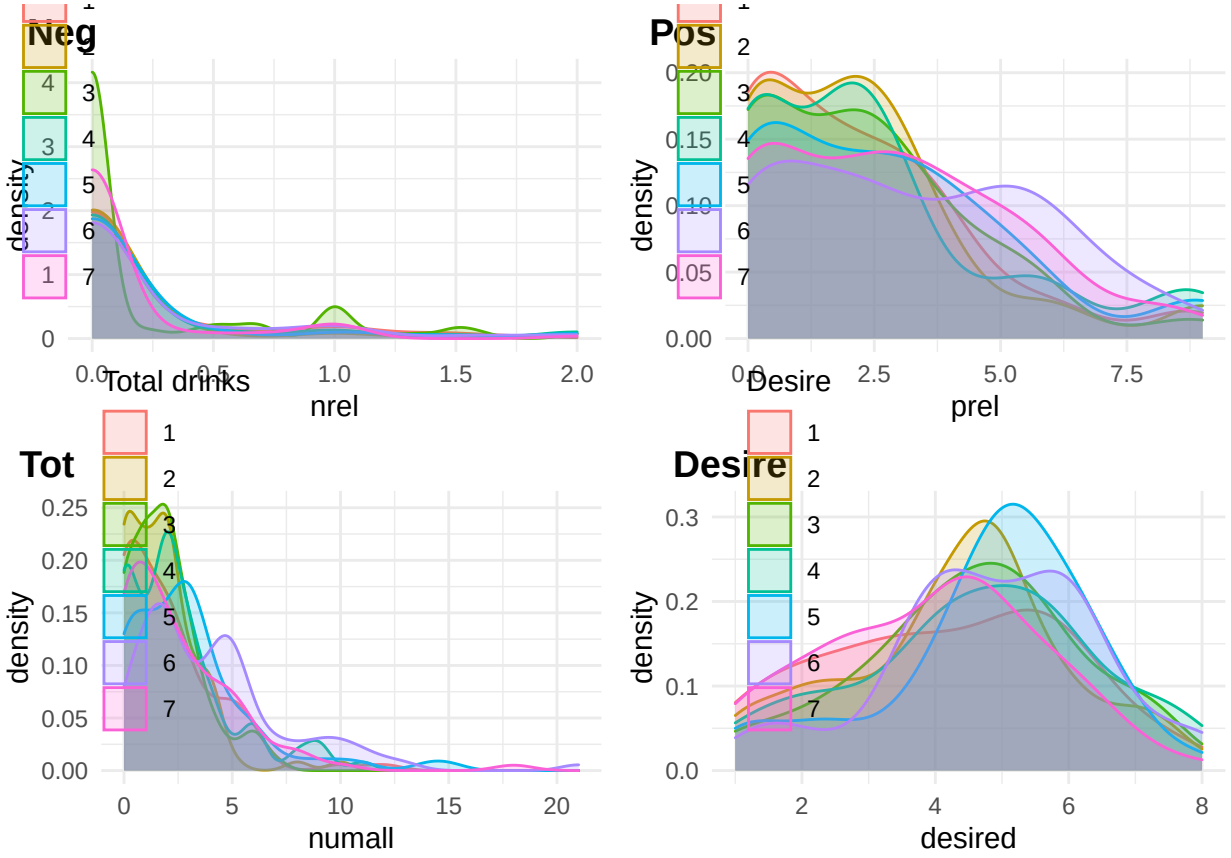

```
library(ggpubr)
neg_plot <- ggplot(data = drinks,
                   aes(x = nrel,
                       fill = as.factor(dayweek),
                       colour = as.factor(dayweek))
                       ) +
              geom_density(alpha = 0.2) +
              theme_minimal() +
              theme(legend.position = c(0.15, 0.8)) +
              guides(fill = guide_legend(title = "Neg romantic event"),
                     colour = guide_legend(title = "Neg romantic event")) +
                xlim(0, 2)

pos_plot <- ggplot(data = drinks,
                   aes(x = prel,
                       fill = as.factor(dayweek),
                       colour = as.factor(dayweek))
                       ) +
              geom_density(alpha = 0.2) +
              theme_minimal() +
              theme(legend.position = c(0.15, 0.8)) +
              guides(fill = guide_legend(title = "Pos romantic event"),
                     colour = guide_legend(title = "Pos romantic event"))
```

```r
tot_plot <- ggplot(data = drinks,
                   aes(x = numall,
                       fill = as.factor(dayweek),
                       colour = as.factor(dayweek))
                       ) +
            geom_density(alpha = 0.2) +
            theme_minimal() +
            theme(legend.position = c(0.15, 0.8)) +
            guides(fill = guide_legend(title = "Total drinks"),
                   colour = guide_legend(title = "Total drinks"))

sest_plot <- ggplot(data = drinks,
                   aes(x = desired,
                       fill = as.factor(dayweek),
                       colour = as.factor(dayweek))
                       ) +
            geom_density(alpha = 0.2) +
            theme_minimal() +
            theme(legend.position = c(0.15, 0.8)) +
            guides(fill = guide_legend(title = "Desire"),
                   colour = guide_legend(title = "Desire"))

ggarrange(NULL, NULL, neg_plot, pos_plot, NULL, NULL, tot_plot, sest_plot,
          labels = c("Neg", "Pos", "", "", "Tot", "Desire", "", ""),
          ncol = 2, nrow = 4, heights = c(0.1, 1, 0.1, 1))
```

```
## Warning: Removed 30 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 3 rows containing non-finite values (stat_density).
```




```
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
```

```
stargazer(as.data.frame(drinks), header = FALSE, type = "latex")
```

'Fill this in: What do you learn?'

2.2 Hypothesis One

(2 points) The researchers hypothesize that negative interactions with romantic partners would be associated with alcohol consumption and an increased desire to drink. Using appropriate models, evaluate the evidence that negative relationship interactions are associated with higher alcohol consumption and an increased desire to drink.

'Fill this in: What do you learn?'

2.3 Hypothesis Two

(1 point) The researchers hypothesize that the relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem. Conduct an analysis

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
id	623	75.888	49.901	1	160
studyday	623	4.000	2.002	1	7
dayweek	623	4.000	2.002	1	7
numall	622	2.524	2.660	0	21
nrel	623	0.359	0.940	0.000	9.000
prel	623	2.583	2.388	0.000	9.000
negevent	623	0.441	0.387	0.000	2.377
posevent	623	1.048	0.646	0.000	3.883
gender	623	1.562	0.497	1	2
rosn	623	3.436	0.420	2.100	4.000
age	623	34.293	4.512	24.433	42.278
desired	620	4.465	1.694	1.000	8.000
state	620	3.966	0.440	2.333	5.000

to address this hypothesis.

'Fill this in: What do you learn?'