Lab 1, Short Questions

Contents

```r
# install.packages("MASS")
# install.packages("ggplot2")
# install.packages("MASS")
# install.packages("GGally")
library(tidyverse)
library(patchwork)
library(magrittr)
library(MASS)
library(ggplot2)
library(GGally)
library(knitr)
library(gridExtra)

# multinomial regression
library(nnet)

# car pacakge for testing
library(car)

theme_set(theme_minimal())
knitr::opts_chunk$set(tidy.opts = list(width_cutoff = 100), tidy = TRUE)
knitr::opts_chunk$set(message = FALSE)
```

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.*

```
cereal <- read_csv("../data/short-questions/cereal_dillons.csv")
```

1.1 Recode Data

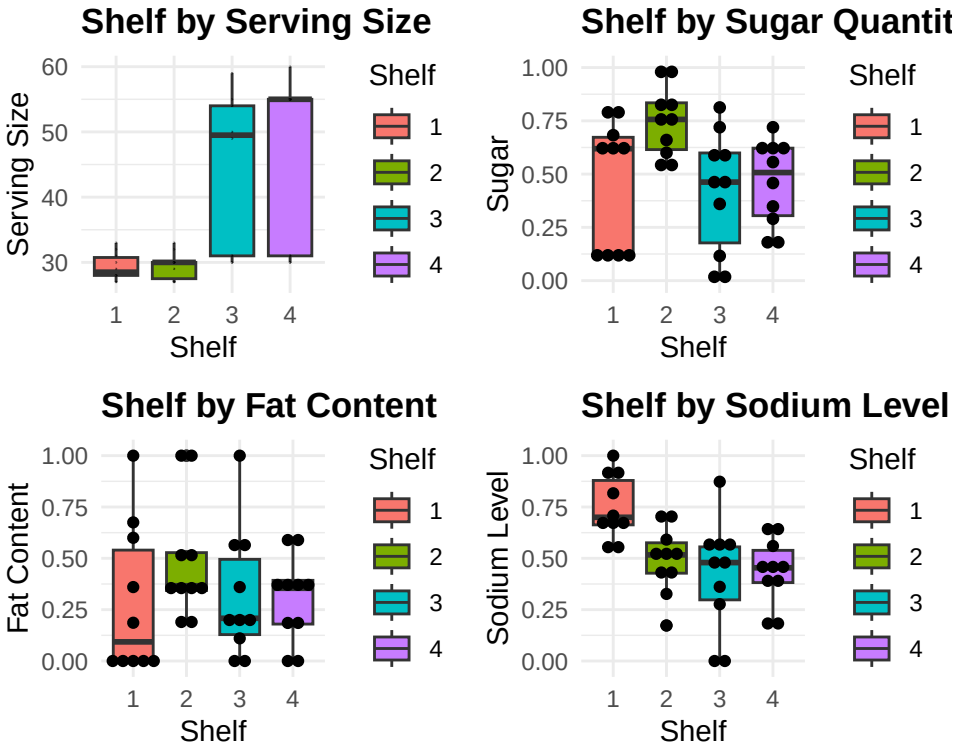
(1 point) The explanatory variables need to be reformatted before proceeding further (sample code is provided in the textbook). First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Second, rescale each variable to be within 0 and 1. 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.

```
# reformatting the explanatory variables
# cereal$Shelf <- factor(cereal$Shelf)

# normalize the explanatory variables
# 1. divide by serving size
# 2. normalized

stand01 <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

cereal2 <- data.frame(
  Shelf = cereal$Shelf,
  size = cereal$size_g,
  sugar = stand01(x = cereal$sugar_g / cereal$size_g),
  fat = stand01(x = cereal$fat_g / cereal$size_g),
  sodium = stand01(x = cereal$sodium_mg / cereal$size_g)
)
summary(cereal2)
```

```
##      Shelf           size           sugar             fat
##  Min.   :1.00   Min.   :27.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.75   1st Qu.:29.75   1st Qu.:0.3339   1st Qu.:0.1582
##  Median :2.50   Median :31.00   Median :0.6000   Median :0.3542
```

```
##  Mean   :2.50   Mean   :37.20   Mean   :0.5209   Mean   :0.3476
##  3rd Qu.:3.25   3rd Qu.:51.00   3rd Qu.:0.7200   3rd Qu.:0.5400
##  Max.   :4.00   Max.   :60.00   Max.   :1.0000   Max.   :1.0000
##      sodium
##  Min.   :0.0000
##  1st Qu.:0.4200
##  Median :0.5354
##  Mean   :0.5240
##  3rd Qu.:0.6696
##  Max.   :1.0000
```

```r
p1 <- cereal2 %>%
  mutate(Shelf = as.factor(Shelf)) %>%
  ggplot(aes(Shelf, size)) +
  geom_boxplot(aes(fill = Shelf)) +
  geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.05) +
  # coord_flip() +
  ggtitle("Shelf by Serving Size") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab("Serving Size") +
  xlab("Shelf")

p2 <- cereal2 %>%
  mutate(Shelf = as.factor(Shelf)) %>%
  ggplot(aes(Shelf, sugar)) +
  geom_boxplot(aes(fill = Shelf)) +
  geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.05) +
  # coord_flip() +
  ggtitle("Shelf by Sugar Quantity") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab("Sugar") +
  xlab("Shelf")

p3 <- cereal2 %>%
  mutate(Shelf = as.factor(Shelf)) %>%
  ggplot(aes(Shelf, fat)) +
  geom_boxplot(aes(fill = Shelf)) +
  geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.05) +
  # coord_flip() +
  ggtitle("Shelf by Fat Content") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab("Fat Content") +
  xlab("Shelf")

p4 <- cereal2 %>%
  mutate(Shelf = as.factor(Shelf)) %>%
  ggplot(aes(Shelf, sodium)) +
  geom_boxplot(aes(fill = Shelf)) +
  geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.05) +
  # coord_flip() +
  ggtitle("Shelf by Sodium Level") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab("Sodium Level") +
  xlab("Shelf")


grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



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

- For the sugar variable, we observe that:

  1) Of the four shelves, Shelf 2 has the highest mean sugar but also has the smallest variance (consistency within the shelf);

  2) Shelf 3 has the lowest mean sugar, but also has relatively large variance among the items on the shelf.

- For the fat variable, we observe that

  1) Of the four shelves, Shelf 2 has the highest mean fat but has relatively smaller variance, similar observation as for the sugar variable;

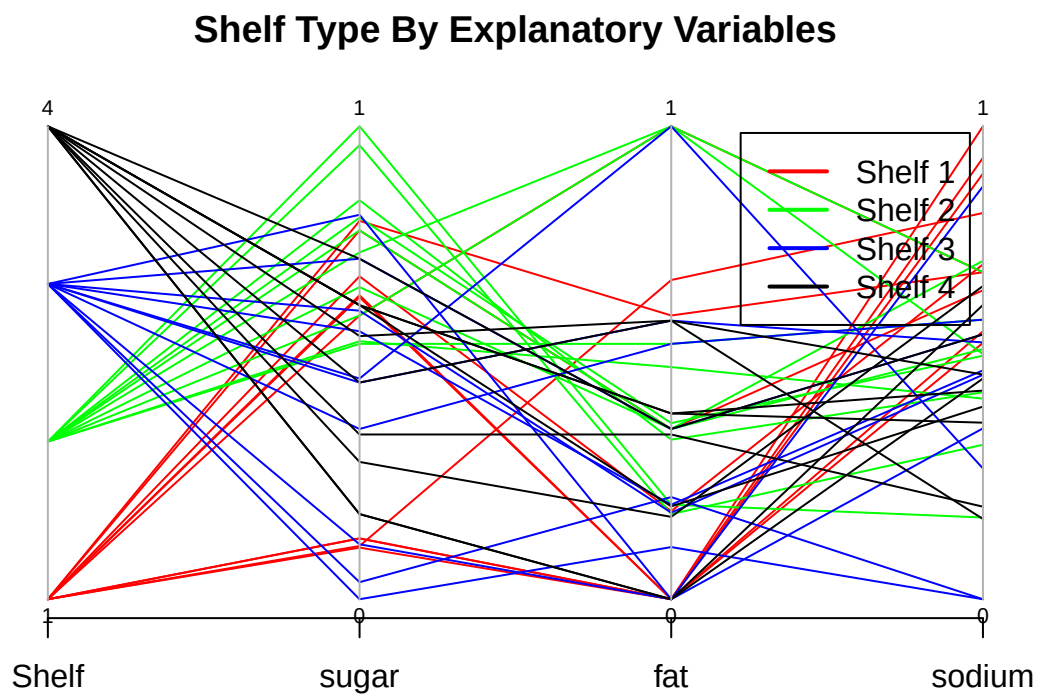
  2) Shelf 1 and 3 have lower fat, but relatively large variance

- For the sodium variable, we observe that

  1) Of the four shelves, Shelf 2, 3 and 4 have similar mean sodium, but Shelf 2 and 4 have relatively smaller variance.

  2) Shelf 1 has the highest sodium'

```r
# for ease of visualization, Adding shelf here.
cols <- c("red", "green", "blue", "black")
parcoord(cereal2[c("Shelf", "sugar", "fat", "sodium")],
  col = cols[cereal2$Shelf], var.label = TRUE,
  main = "Shelf Type By Explanatory Variables"
)
legend("topright",
  legend = c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
  lwd = 2, col = cols, inset = 0.05
)
```



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

For Shelf 2 (green), sugar contents are generally higher, follow by Shelf 1 (red).

Shelf 3 (Blue) has the biggest variance in all categories.

In term of sodium content, Shelf 1 (red) generally shows the highest level.

Shelf 4 (Black), also shows a lower trend on fat content compared to all other Shelves.'

Fill in: Do content differences exist between the shelves?

Yes based on the box plot, there are a clear visualization differences between the mean and variance for each individual contents among the shelves.

For example, Shelf 2 generally higher sugar/fat content per serving size, which could extract a lot more sales among younger children.'

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?
>
> Since argument can be made for both nominal vs ordinal scale for the Shelf level, we decided that taking account to ordinality scale is not desireable under this setting.
>
> The ordinal data referes to the location of the shelves. Since the shelf system mentioned in this problem is purely due to height. Shelf 1 is at the lowest height, follows by Shelf 2, Shelf 3, and Shelf 4. However, this ordering does not necssary related to to the expected strategical importance from sales perspective.
>
> Along with the graphs, we believe that a natural progression order $1 < 2 < 3 < 4$ is not presence making the shelves placement nominal.'

```
model_cereal_shelves_linear <- multinom(
  formula = factor(Shelf) ~ sugar
    + fat
    + sodium,
  data = cereal2
)
```

```
## # 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
```

```
summary(model_cereal_shelves_linear)
```

```
## Call:
## multinom(formula = factor(Shelf) ~ sugar + fat + sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar        fat    sodium
## 2    6.900708   2.693071  4.0647092 -17.49373
## 3   21.680680 -12.216442 -0.5571273 -24.97850
## 4   21.288343 -11.393710 -0.8701180 -24.67385
```

```
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 2    6.487408 5.051689 2.307250 7.097098
## 3    7.450885 4.887954 2.414963 8.080261
## 4    7.435125 4.871338 2.405710 8.062295
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

```
model_cereal_shelves_cross <- multinom(
  formula = factor(Shelf) ~ sugar
    + fat
    + sodium
    + sugar:fat
    + sugar:sodium
    + fat:sodium
    + sugar:fat:sodium,
  data = cereal2
)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 36.170336
## iter  20 value 31.166546
## iter  30 value 29.963705
## iter  40 value 28.414027
## iter  50 value 27.891712
## iter  60 value 27.763967
## iter  70 value 27.622579
## iter  80 value 27.438263
## iter  90 value 27.015534
## iter 100 value 26.772481
## final  value 26.772481
## stopped after 100 iterations
```

```
summary(model_cereal_shelves_cross)
```

```
## Call:
## multinom(formula = factor(Shelf) ~ sugar + fat + sodium + sugar:fat +
##     sugar:sodium + fat:sodium + sugar:fat:sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar       fat     sodium sugar:fat sugar:sodium fat:sodium
## 2   -4.563627   8.944868 22.063003   1.030077  35.60873   -12.250084  -23.75955
## 3   24.498320 -22.248456 35.981865 -27.899087 -17.12487    13.253103  -59.54150
## 4   27.246742 -21.852777  7.298799 -29.106797  41.08251     2.887805  -30.85250
##   sugar:fat:sodium
## 2        -55.88455
## 3         37.71571
## 4        -22.59552
##
## Std. Errors:
##   (Intercept)    sugar       fat   sodium sugar:fat sugar:sodium fat:sodium
## 2    25.21113 29.72894  96.57821 27.29915  135.1117     31.98647   116.0776
## 3    22.83750 25.81043 101.17670 24.61166  150.1228     26.89827   138.0237
## 4    22.80359 26.00692 100.83444 24.51538  150.6750     28.86631   138.5448
##   sugar:fat:sodium
```

```
## 2         158.8091
## 3         212.2222
## 4         217.3953
##
## Residual Deviance: 53.54496
## AIC: 101.545
```

```
lrt_cereal_main_effects <- Anova(model_cereal_shelves_linear, test = "LR")
lrt_cereal_main_effects
```

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

For the first model with a linear effect, we are using a LRT test for each individual explanatory variable.

The resulted p-values for sugar and sodium contents are both lower than alpha = 0.001, or at 99.99% confidence interval and visualized by three asterisk. This suggest that sugar and sodium content are statistically significant to the model.

The fat content estimated parameter resulted in a p-value of 0.1522, which is higher than all proposed alpha value. Thus, fat content is not statistically significant to the model.

```
lrt_cereal_quadratic_effects <- anova(model_cereal_shelves_linear,
  model_cereal_shelves_cross,
  test = "Chisq"
)
lrt_cereal_quadratic_effects
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: factor(Shelf)
##                                                                                 Model
## 1                                                                  sugar + fat + sodium
## 2 sugar + fat + sodium + sugar:fat + sugar:sodium + fat:sodium + sugar:fat:sodium
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1       108   67.19028
## 2        96   53.54496 1 vs 2    12 13.64531 0.3239288
```

For the second model with quadratic effect. We are using the LRT for a Null hypothesis that the interaction terms among the explanatory variable have estimated parameters of 0, and an alternate hypothesis that they are not. The test resulted in a p-value of 0.3239 from a chi-squared distribution.

Since the p-value is greater than the alpha = 0.05 or at 95% confidence level, the interaction terms are not statistically significant.

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.

```
# create a data set with standardized value
new.data <- data.frame(
  "Cereal" = "Apple Jack",
  "size_g" = 28,
  "sugar_g" = 12,
  "fat_g" = 0.5,
  "sodium_mg" = 130
)

stand02 <- function(new_data, original_data, col_names) {
  for (col_name in col_names) {
    min_val <- min(original_data[col_name] / original_data["size_g"])
    max_val <- max(original_data[col_name] / original_data["size_g"])

    new_col_name <- substring(col_name, 0, unlist(gregexpr('_', col_name))[1] - 1)
    new_data[new_col_name] <- (new_data[col_name] / new_data["size_g"] - min_val) / (max_val -
  }
  return(new_data)
}

new.data <- stand02(new.data, cereal, c("sugar_g", "fat_g", "sodium_mg"))

# get model prediction for the given inputs
aj_shelf_probs <- predict(model_cereal_shelves_linear,
  newdata = new.data,
  type = "probs" # could use type = "class"
)

# aj_shelf_probs
data.frame(pi.hat = round(aj_shelf_probs, 3) * 100)
```

```
##   pi.hat
## 1    5.3
## 2   47.2
## 3   20.0
## 4   27.4
```

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

Based on the outputs, Shelf 2 has the highest probability of 47.2% follows by Shelf 4 with a probability of 27.4%.

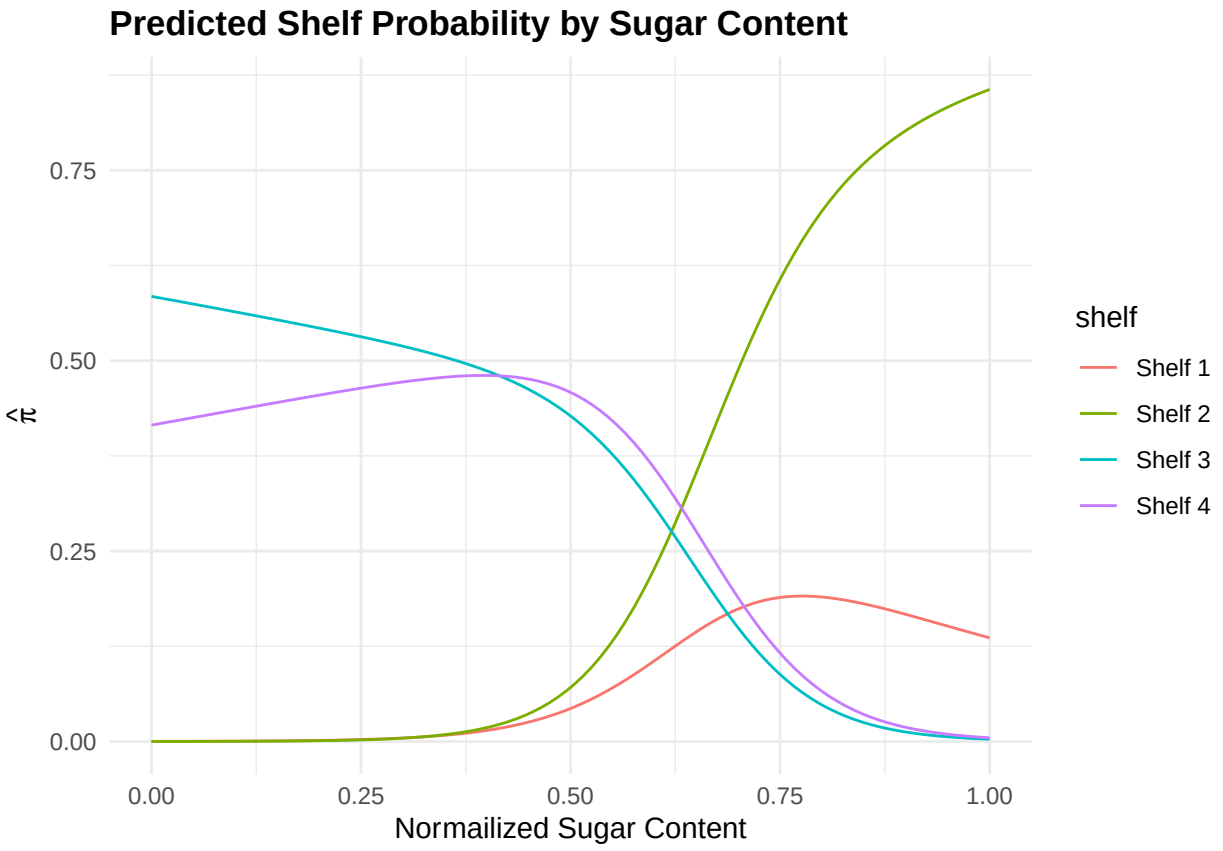
Since Shelf 2 has the highest probability, Apple Jack is predicted to be placed on Shelf 2.

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.

```
# stats taken from summary(cereal2) statistics from earlier

min_sugar <- 0
max_sugar <- 1

mean_fat <- mean(cereal2$fat)
mean_sodium <- mean(cereal2$sodium)

# get prob for every single suger level s, holding fat and sodium constant
shelfs <- function(s) {
  new.data <- data.frame(
    "sugar" = s,
    "fat" = mean_fat,
    "sodium" = mean_sodium
  )

  pred.prob <- predict(model_cereal_shelves_linear,
    newdata = new.data, type = "probs"
  )

  outputs <- data.frame(pi.hat = round(pred.prob, 10))
  return(outputs)
}


model.preds <- data.frame(
  x = seq(0, 1, length.out = 100),
  values = c(
    shelfs(seq(0, 1, length.out = 100))$pi.hat.1,
    shelfs(seq(0, 1, length.out = 100))$pi.hat.2,
    shelfs(seq(0, 1, length.out = 100))$pi.hat.3,
    shelfs(seq(0, 1, length.out = 100))$pi.hat.4
  ),
  shelf = rep(c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
    each = 100
  )
)

ggplot(
  model.preds, # Draw ggplot2 plot
  aes(x, values, col = shelf)
) +
```

```
  geom_line() +
  ggtitle("Predicted Shelf Probability by Sugar Content") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab(expression(hat(pi))) +
  xlab("Normailized Sugar Content")
```



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

Lower sugar contents cereal has higher probability to be place on Shelf 3 and 4. Higher sugar contents have higher probability of being place on Shelf 2.

There seem to be a low probability of cereals being placed on Shelf 1 despite of sugar content, however the trend does follow that of Shelf 2. This make sense because Shelf 1 and 2 are have a lower height than Shelf 3 and 4.

High sugar contents cereal generally attract younger demographics than low sugar contents cereal, which older demographics might deem as healthier options.'

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.

```
# get the standard deviations
sd.cereal <- apply(X = cereal2[c("sugar", "fat", "sodium")], MARGIN = 2, FUN = sd)
c.value <- round(sd.cereal, 4)

# get the estimated parameters for each shelf level
beta.hat2 <- coefficients(model_cereal_shelves_linear)[1, 2:4]
beta.hat3 <- coefficients(model_cereal_shelves_linear)[2, 2:4]
beta.hat4 <- coefficients(model_cereal_shelves_linear)[3, 2:4]

# using confint to compute CI for each explanatory variable
conf.beta <- confint(object = model_cereal_shelves_linear, level = 0.95)

# Odds for j = 2 vs j = 1 (Shelf 2 vs Shelf 1)
OR_2_1 <- round(exp(c.value * beta.hat2), 4)
OR_CI_2_1 <- round(exp((c.value * conf.beta[2:4, 1:2, 1])), 4)

# Odds for j = 3 vs j = 1 (Shelf 3 vs Shelf 1)
OR_3_1 <- round(exp(c.value * beta.hat3), 4)
OR_CI_3_1 <- round(exp((c.value * conf.beta[2:4, 1:2, 2])), 4)

# Odds for j = 4 vs j = 1 (Shelf 4 vs Shelf 1)
OR_4_1 <- round(exp(c.value * beta.hat4), 4)
OR_CI_4_1 <- round(exp((c.value * conf.beta[2:4, 1:2, 3])), 4)

odds_ratios <- cbind(c.value, cbind(OR_2_1, OR_CI_2_1), cbind(OR_3_1, OR_CI_3_1), cbind(OR_4_1
odds_ratios
```

```
##        c.value OR_2_1  2.5 %  97.5 % OR_3_1  2.5 % 97.5 % OR_4_1  2.5 % 97.5 %
## sugar   0.2692 2.0647 0.1436 29.6766 0.0373 0.0028 0.4918 0.0466 0.0036 0.6084
## fat     0.2990 3.3715 0.8722 13.0327 0.8466 0.2056 3.4857 0.7709 0.1883 3.1571
## sodium  0.2298 0.0180 0.0007  0.4389 0.0032 0.0001 0.1224 0.0034 0.0001 0.1302
```

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

Considering only the **statistically significant results at 95%** (whose confidence interval doesn't contain 1 in it), we have the following findings:

For sugar:

The odds of a cereal being place on Shelf 3 is 0.04 times (between 0.0028 to 0.4918 times at 95% confidence level) vs being place on Shelf 1, for a 0.2692 (one sd) increase in sugar content holding all other variable constant.

The odds of a cereal being place on Shelf 4 is 0.05 times (between 0.0036 to 0.6084 times at 95% confidence level) vs being place on Shelf 1, for a 0.2692 (one sd) increase in sugar content holding all other variable constant.

This agrees with our observation that, higher sugar cereals tends to have higher probability / odds ratio to be on shelf 1 or 2, vs shelf 3 or 4.

For fat:

Fat level didn't seem to provide any significant association across different shelves, so we are skipping the interpretation here (the odds ratio would have similar interpretation as sugar).

For sodium:

The odds of a cereal being place on Shelf 2 is 0.018 times (between 0.0007 to 0.4389 times at 95% confidence level) vs being place on Shelf 1, for a 0.2298 (one sd) increase in sodium content holding all other variable constant.

The odds of a cereal being place on Shelf 3 is 0.0032 times (between 0.0001 to 0.1224 times at 95% confidence level) vs being place on Shelf 1, for a 0.2298 (one sd) increase in sodium content holding all other variable constant.

The odds of a cereal being place on Shelf 4 is 0.0034 times (between 0.0001 to 0.1302 times at 95% confidence level) vs being place on Shelf 1, for a 0.2298 (one sd) increase in sodium content holding all other variable constant.

Overall, it seem shelf 1 have very high sodium cereals, which is inline with our observation - thus increasing the level of sodium would make it very likely to be on shelf 1 rather than other shelves.

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.*

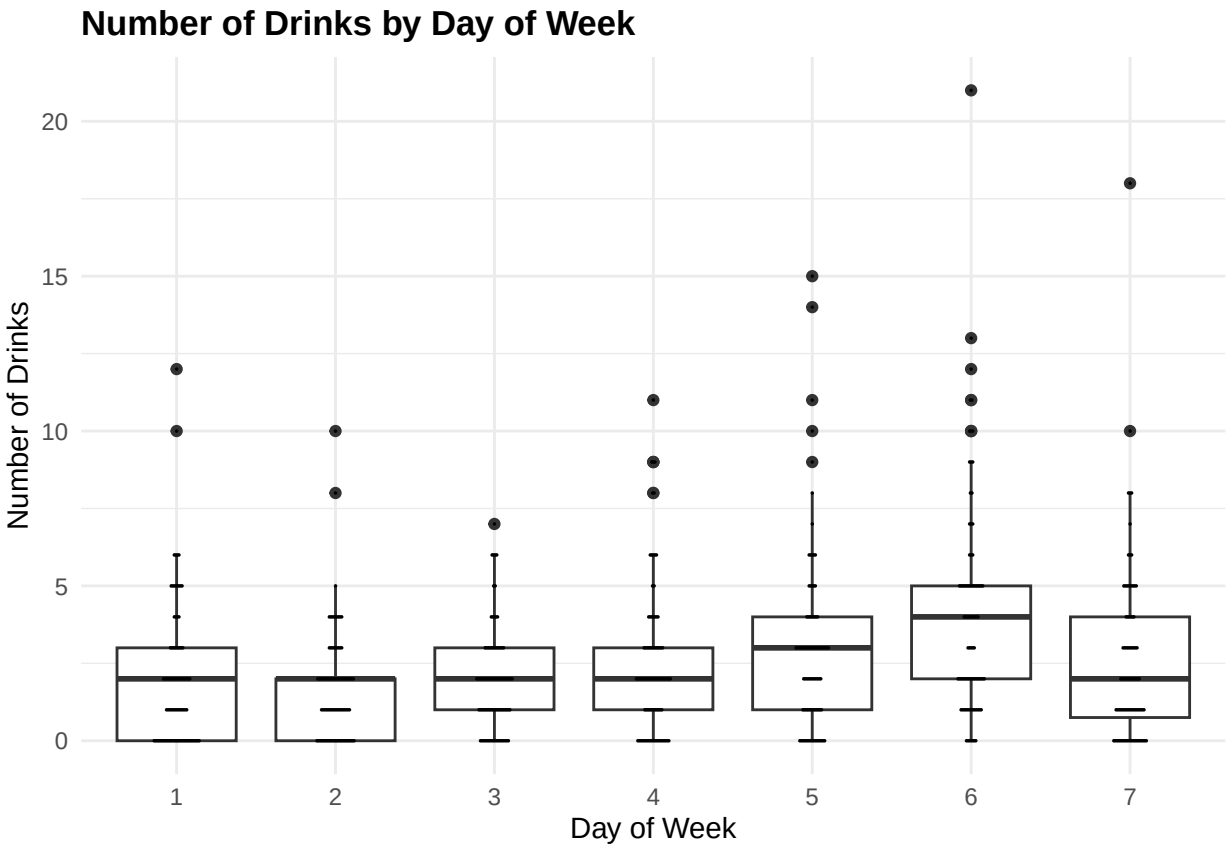
```
drinks <- read_csv("../data/short-questions/DeHartSimplified.csv")

summary(drinks)
```

```
##        id            studyday    dayweek      numall            nrel
##  Min.   :  1.00   Min.   :1   Min.   :1   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 33.00   1st Qu.:2   1st Qu.:2   1st Qu.: 1.000   1st Qu.:0.000
##  Median : 60.00   Median :4   Median :4   Median : 2.000   Median :0.000
##  Mean   : 75.89   Mean   :4   Mean   :4   Mean   : 2.524   Mean   :0.359
##  3rd Qu.:123.00   3rd Qu.:6   3rd Qu.:6   3rd Qu.: 3.750   3rd Qu.:0.000
##  Max.   :160.00   Max.   :7   Max.   :7   Max.   :21.000   Max.   :9.000
##                                           NA's   :1
##       prel           negevent         posevent         gender
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :1.000
##  1st Qu.:0.4167   1st Qu.:0.1583   1st Qu.:0.600   1st Qu.:1.000
##  Median :2.0000   Median :0.3500   Median :0.950   Median :2.000
##  Mean   :2.5830   Mean   :0.4414   Mean   :1.048   Mean   :1.562
##  3rd Qu.:4.0000   3rd Qu.:0.6292   3rd Qu.:1.378   3rd Qu.:2.000
##  Max.   :9.0000   Max.   :2.3767   Max.   :3.883   Max.   :2.000
##
##       rosn            age           desired          state
##  Min.   :2.100   Min.   :24.43   Min.   :1.000   Min.   :2.333
##  1st Qu.:3.200   1st Qu.:30.53   1st Qu.:3.333   1st Qu.:3.667
##  Median :3.500   Median :34.57   Median :4.667   Median :4.000
##  Mean   :3.436   Mean   :34.29   Mean   :4.465   Mean   :3.966
##  3rd Qu.:3.800   3rd Qu.:38.19   3rd Qu.:5.667   3rd Qu.:4.222
##  Max.   :4.000   Max.   :42.28   Max.   :8.000   Max.   :5.000
##                                  NA's   :3       NA's   :3
```

2.1 EDA

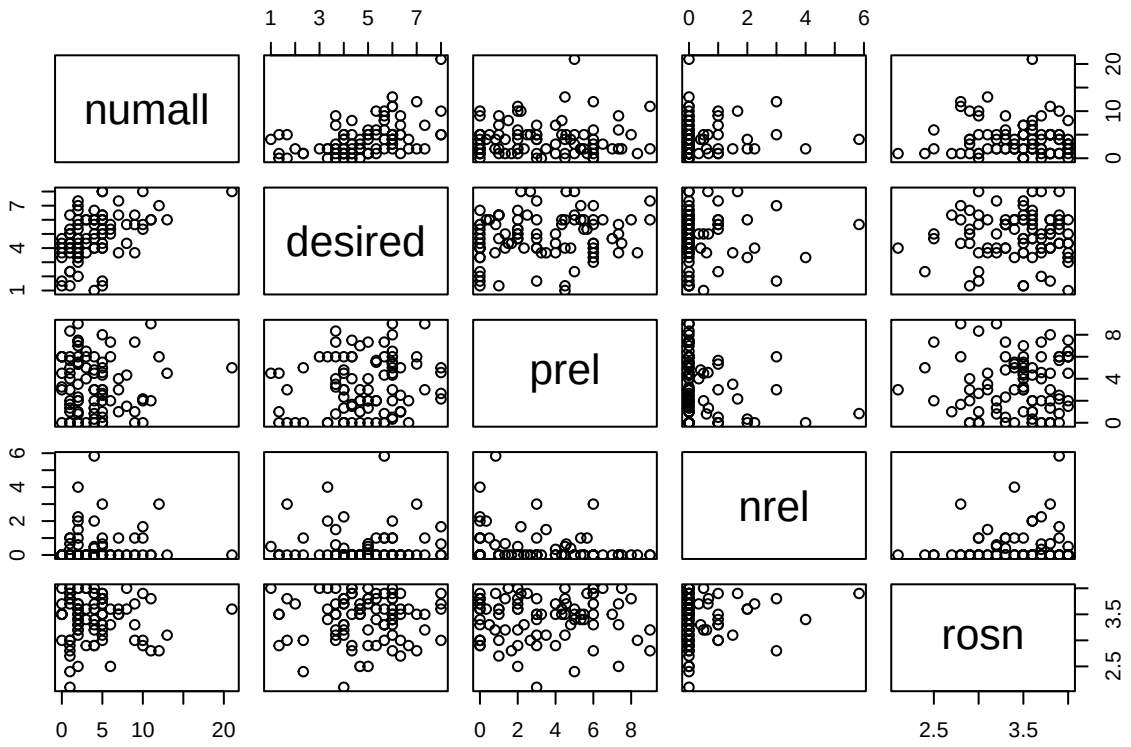
(2 points) Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. Address the reasons for limiting the study to observations from only one day.

```
drinks %>%
  dplyr::mutate(dow_factor = factor(dayweek)) %>%
  dplyr::select(dow_factor, numall) %>%
  drop_na() %>%
  ggplot(aes(dow_factor, numall)) +
  geom_boxplot(aes(dow_factor)) +
  geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.05) +
  ggtitle("Number of Drinks by Day of Week") +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ylab("Number of Drinks") +
  xlab("Day of Week")
```



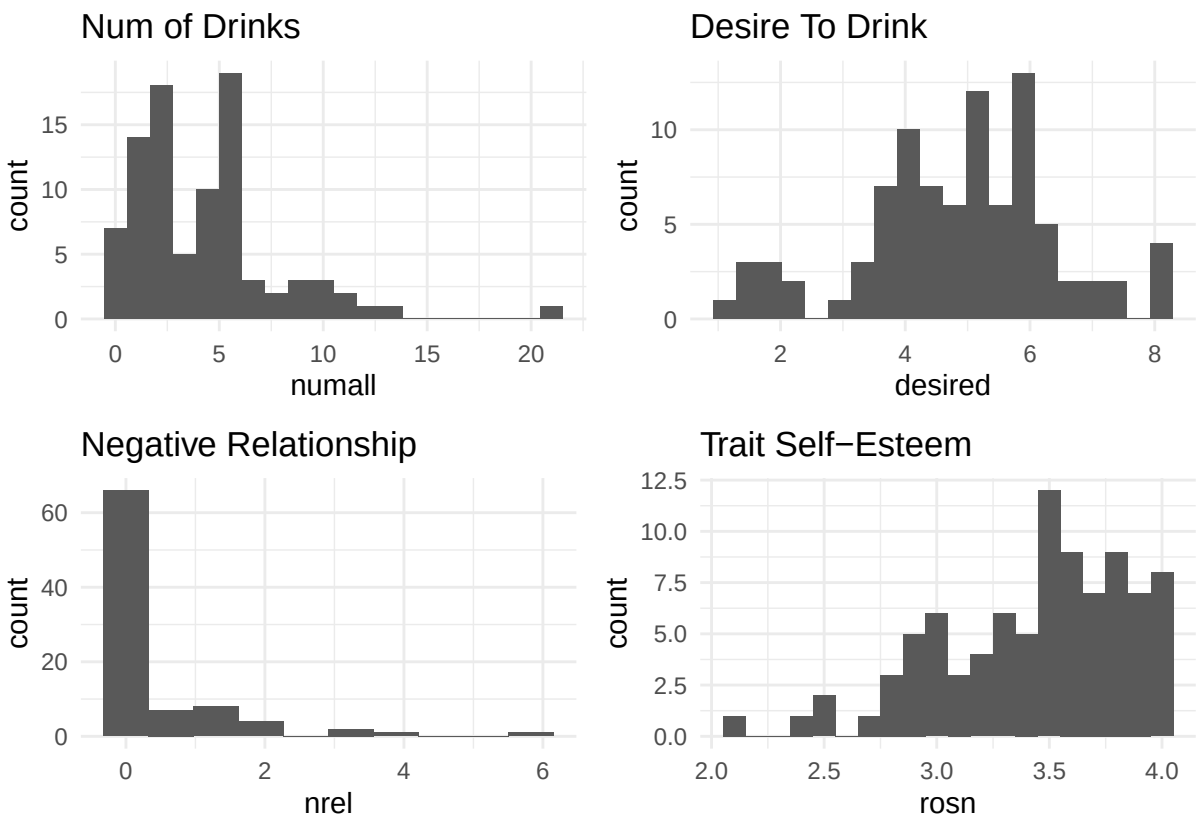
- We can observe that the day of the week seem to have association with the number of drinks consumed, and Saturday tends to have the highest number of drinks in our sample.
- We will limit the study to observations from only Saturday, as the researchers stated that they are interested in the relationship between alcohol consumption and negative relationship interactions, and we need to remove the confounding effect of day of the week.

```
sat_drinks <- drinks %>%
  dplyr::filter(dayweek == 6) %>%
  dplyr::select(numall, desired, prel, nrel, rosn) %>%
  drop_na()

pairs(sat_drinks)
```



- We spot "numall" (number of drinks) appear to have a positive correlation with "desired", which is expected as the more one desire, the more one might end up drinking.
- There doesn't seem to have any evidence of perfect co-linearity based on the plots.

```r
h1 <- sat_drinks %>%
  ggplot(aes(numall)) +
  geom_histogram(bins = 20) +
  ggtitle("Num of Drinks")

h2 <- sat_drinks %>%
  ggplot(aes(desired)) +
  geom_histogram(bins = 20) +
  ggtitle("Desire To Drink")

h3 <- sat_drinks %>%
  ggplot(aes(nrel)) +
  geom_histogram(bins = 10) +
  ggtitle("Negative Relationship")


h4 <- sat_drinks %>%
  ggplot(aes(rosn)) +
  geom_histogram(bins = 20) +
  ggtitle("Trait Self-Esteem")

(h1 | h2) / (h3 | h4)
```

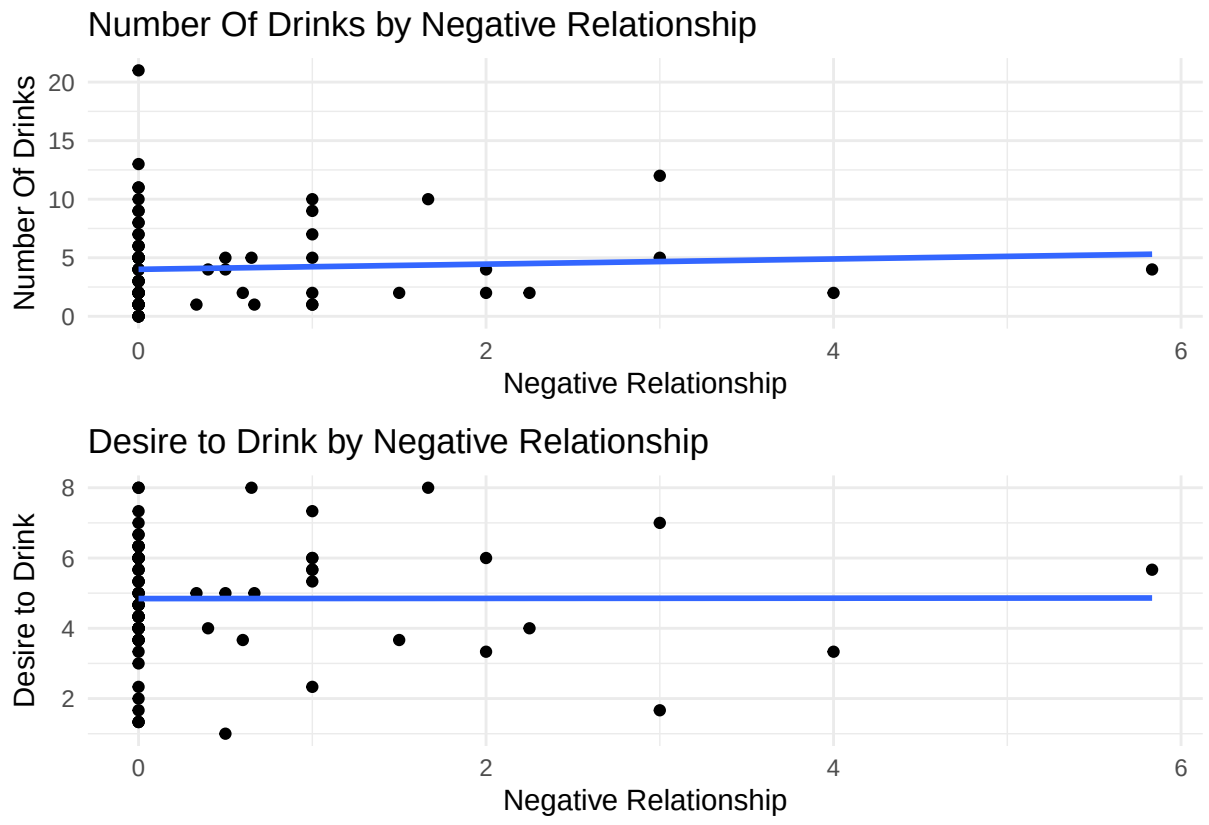


```r
sat_drinks %>%
  mutate(rounded_nrel = round(nrel, 0)) %>%
  count(rounded_nrel) %>%
  mutate(prop = round(prop.table(n), 2)) %>%
  kable(col.names = c("Negative Relationship", "Count", "Proportion"))
```

Negative Relationship	Count	Proportion
0	70	0.79
1	10	0.11
2	5	0.06
3	2	0.02
4	1	0.01
6	1	0.01

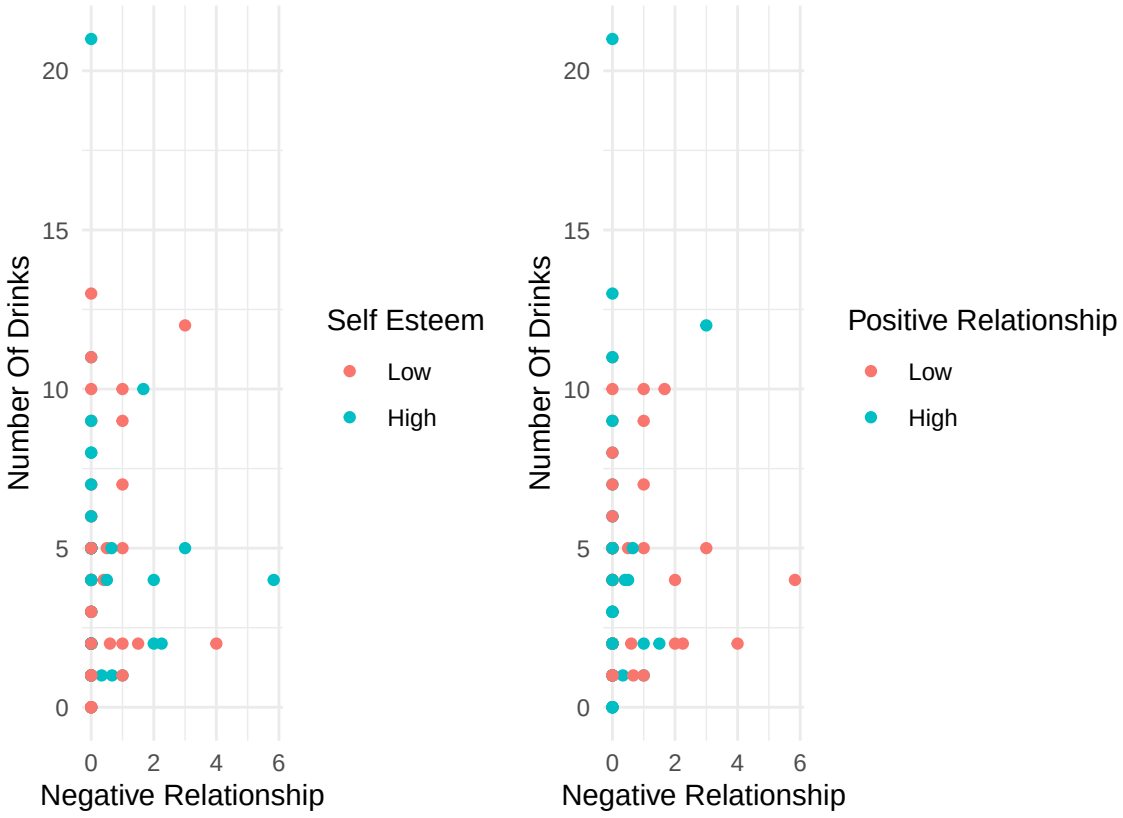
- The number of drinks is right-skewed with a lower bound of 0, which is expected.
- The desire to drink seems balanced within our sample.
- Negative Relationship is heavily right-skewed, with 70, or 79% of the observations with 0 negative relationship. This might be a problem given the limited sample size. We might not be able to properly test this hypothesis.
- Trait self-esteem is left-skewed with a cap of 4.0 and min of 2.0.

```r
p2_1 <- sat_drinks %>%
  ggplot(aes(nrel, numall)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Number Of Drinks by Negative Relationship") +
  ylab("Number Of Drinks") +
  xlab("Negative Relationship")

p2_2 <- sat_drinks %>%
  ggplot(aes(nrel, desired)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Desire to Drink by Negative Relationship") +
  ylab("Desire to Drink") +
  xlab("Negative Relationship")

p2_1 / p2_2
```



Given the limited number of observations that have more than 1 negative relationship reading, it's very hard to see a trend via the bi-variant visualization.

```r
sat_drinks <- sat_drinks %>%
  mutate(high_self_esteem = rosn > median(rosn))

p3_1 <- sat_drinks %>%
  ggplot(aes(x = nrel, y = numall, color = high_self_esteem)) +
  geom_point() +
  coord_quickmap() +
  ylab("Number Of Drinks") +
  xlab("Negative Relationship") +
  scale_color_discrete(name = "Self Esteem", labels = c("Low", "High"))


p3_2 <- sat_drinks %>%
  mutate(high_positive_relationship = prel > median(prel)) %>%
  ggplot(aes(x = nrel, y = numall, color = high_positive_relationship)) +
  geom_point() +
  coord_quickmap()+
  ylab("Number Of Drinks") +
  xlab("Negative Relationship") +
  scale_color_discrete(name = "Positive Relationship", labels = c("Low", "High"))

p3_1 | p3_2
```



- From the first plot, we can see that at a given negative relationship event, we tend to observe more number of drinks for low self-esteem observations (in red) than high self-esteem observations.

  This is consistent with the assumption that the researcher has.

- From the first plot, we can see that at a given negative relationship event, we tend to observe more number of drinks for observations with less positive relationship event (in red) than high positive relationship observations.

  This make us wonder if including the negative relationship itself is enough to predict the alcohol consumption, given the negative relationship effects could well be offset by positive relationship effects.

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.

```
mod_nrel <-
  sat_drinks %>% glm(
    formula = numall ~ nrel,
    family = poisson(link = log)
  )

summary(mod_nrel)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = log), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8337  -1.3211  -0.5305   0.4733   5.9597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39003    0.05715  24.320   <2e-16 ***
## nrel         0.04971    0.05076   0.979    0.328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.43  on 87  degrees of freedom
## AIC: 508.83
##
## Number of Fisher Scoring iterations: 5
```

- First, we use the Poisson model to predict the number of drinks using only the negative relationship event as the explanatory variable.
- Based on the output, we failed to reject the null hypothesis that negative relationship has no impact on the number of drinks consumed.
- We believe this is mostly because the lack of data: only 19 observations among a total of 89 have non-zero negative relationship event.

```r
mod_nrel_and_prel <-
  sat_drinks %>% glm(
    formula = numall ~ nrel + prel,
    family = poisson(link = log)
  )

summary(mod_nrel_and_prel)
```

```
##
## Call:
## glm(formula = numall ~ nrel + prel, family = poisson(link = log),
##     data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8732  -1.2742  -0.4526   0.5581   5.9134
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.34901    0.09553  14.121   <2e-16 ***
## nrel         0.05684    0.05235   1.086    0.278
## prel         0.01145    0.02115   0.541    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.14  on 86  degrees of freedom
## AIC: 510.54
##
## Number of Fisher Scoring iterations: 5
```

- Then, we included positive relationship as a second explanatory variable, in an attempt to adjust for the offset emotional effect from positive relationship events on the number of alcohol assumptions.
- Using the Poisson model, we still fail to reject the null hypothesis that negative or positive relationship has no impact on the number of drinks consumed.
- Interestingly, the coefficient on positive relationship interactions is positive, meaning there is actually an predicted increase in alcohol consumption with the increase in positive relationship interactions, so our theory has completely fail in this case.

```r
mod_desired <-
  sat_drinks %>% glm(formula = desired ~ nrel, data = .)

summary(mod_desired)
```

```
##
## Call:
## glm(formula = desired ~ nrel, data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8467  -0.8453   0.1533   1.1547   3.1547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.845267   0.184642  26.241   <2e-16 ***
## nrel        0.002914   0.178607   0.016    0.987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.572294)
##
##     Null deviance: 223.79  on 88  degrees of freedom
## Residual deviance: 223.79  on 87  degrees of freedom
## AIC: 340.64
##
## Number of Fisher Scoring iterations: 2
```

- Lastly, we use a linear regression model to predict the desire to drink, using the negative relationship event as the sole explanatory variable.
- Again, we fail to reject the null hypothesis that negative relationship events have no impact on the desire to drink.

Overall, we believe there is no strong evidence that negative relationship events would impact the number of drinks one would consume, given the data set we have chosen.

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 to address this hypothesis.

```
model.neg.relationship.rosn <- sat_drinks %>%
  glm(
    formula = numall ~ nrel + rosn + nrel:rosn,
    family = poisson(link = "log")
  )
summary(model.neg.relationship.rosn)
```

```
##
## Call:
## glm(formula = numall ~ nrel + rosn + nrel:rosn, family = poisson(link = "log"),
##     data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8324  -1.6025  -0.1471   0.5059   5.9811
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.32343    0.46367   2.854  0.00431 **
## nrel         1.07253    0.45716   2.346  0.01897 *
## rosn         0.01642    0.13403   0.123  0.90248
## nrel:rosn   -0.28731    0.13036  -2.204  0.02752 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 244.30  on 85  degrees of freedom
## AIC: 507.7
##
## Number of Fisher Scoring iterations: 5
```

- Using the Poisson regression model, we reject the null hypothesis that the negative relationship has no impact on drinking.
  Given the coefficient of 1.07, we believe more negative relationship will cause more drinking.
- We also reject the null hypothesis that there is no interactions between negative relationship and trait self-esteem.

  Given the coefficient of -0.29, we believe that for people with high self-esteem, their drinking behavior less likely to be affected by an increase in negative relationship events, compared with people with low self-esteem.

- Our model supported the hypothesis of the researcher.