W271 Lab 1, Short Questions

Michael Denton, Srila Maiti, Olivia Pratt, Emily Robles, Elizabeth Willard

Contents

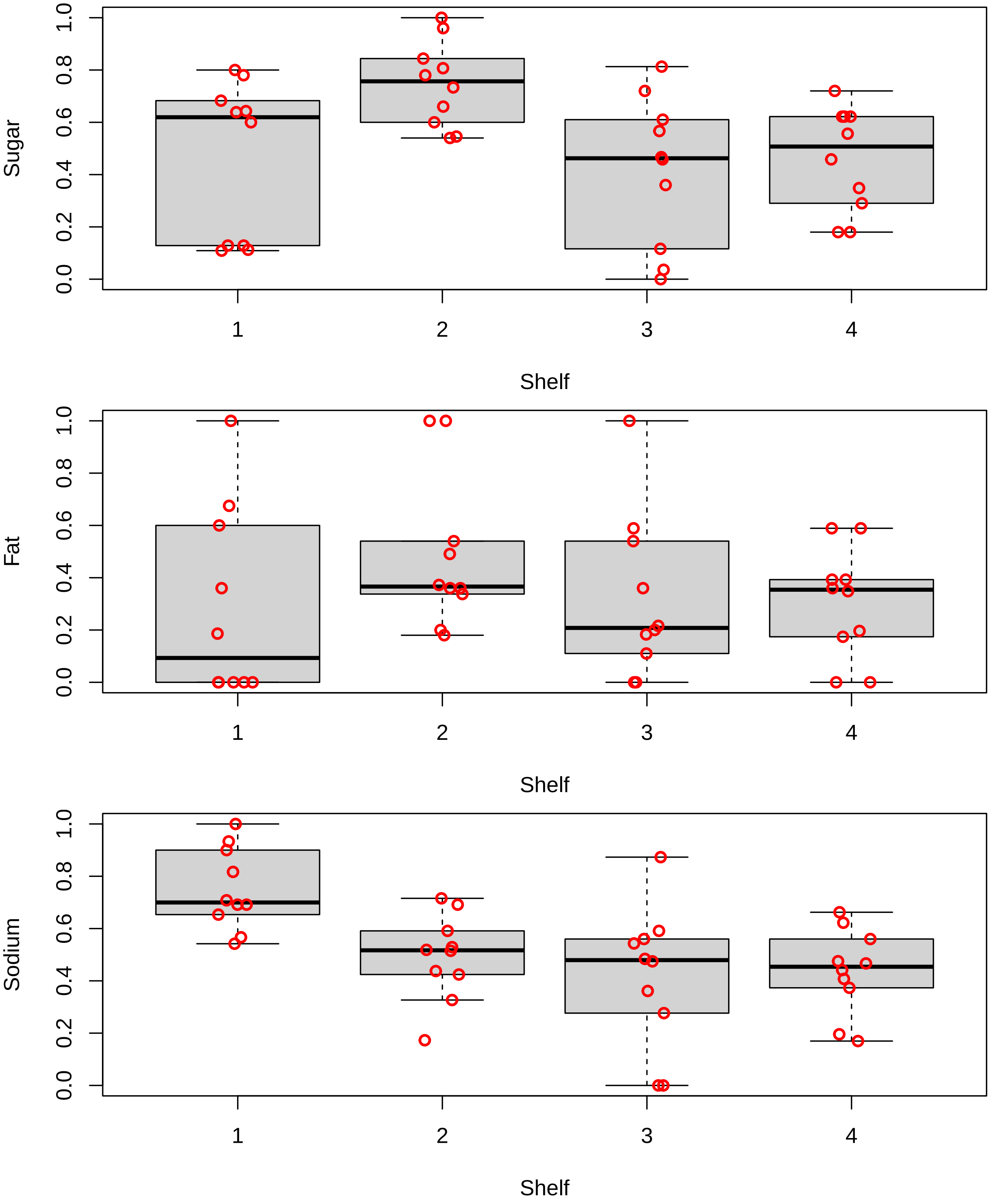
1 Strategic Placement of Products in Grocery Stores (5 points)

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

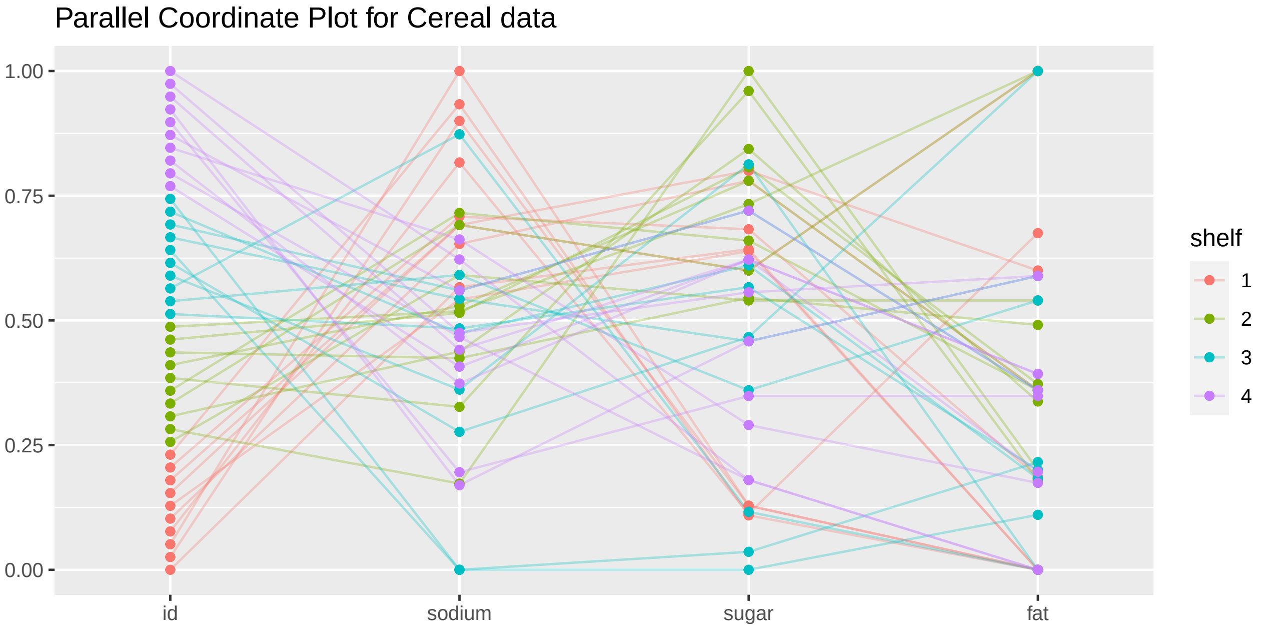
> *In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item—breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular 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.*

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, re scale 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.



The cereals on the second shelf from the bottom generally has much higher suger contents. The second shelf is great real-estate for products, especially ones marketed towards children who are generally shorter than adults. This placement could be used to get kids to ask their parents for more sugary (expensive) cereal. These boxplots also show that cereals with higher sodium tend to be on the first shelf. The fat content of cereals appears to be uniform across the four shelves.



Similar to what we saw in the box plots, there seems to be more high-sugar cereals on the 2nd shelf and more high-sodium cereals on the first shelf.

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).

If there existed a ranking system that determined where a box of cereal was placed on the shelves, we could utilize a ordinal logistic regression model which comes with the benefit of the proportional odds assumption. In this situation however, ordering the data doesn't seem sensible. Common sense would say that having your cereal advertised on the 2nd or 3rd shelves would be ideal, as those are in the eyelines of most people. This data could potentially be ordinal if cereal price was also available, because more expensive or popular brands pay to have their cereal in more visible shelves, but this system does not seem dependent on nutritional content alone.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: 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
## Analysis of Deviance Table (Type II tests)
##
## Response: shelf
##                  LR Chisq Df Pr(>Chisq)
## sugar             19.2525  3  0.0002424 ***
## fat                6.1167  3  0.1060686
## sodium            30.8407  3  9.183e-07 ***
## sugar:fat          3.2309  3  0.3573733
## sugar:sodium       3.0185  3  0.3887844
## fat:sodium         3.1586  3  0.3678151
## sugar:fat:sodium   5.0161  3  0.1706220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This ANOVA test fails to reject the null hypothesis that this model is not improved by interaction terms, because the P values for all of the interaction terms are larger than .05. Thus, we can continue using the model without interaction terms.

1.3 Where do you think Apple Jacks will be placed?

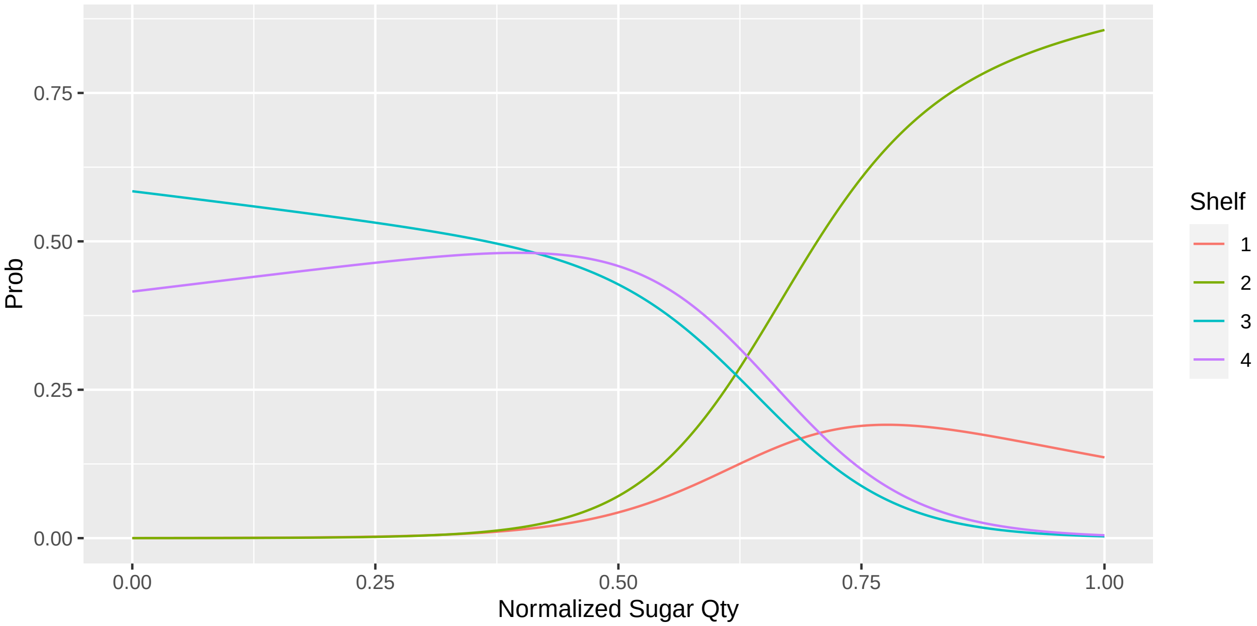
(1 point) Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed toward children. For a serving size of 28 grams, its sugar content is 12 grams, fat content is 0.5 grams, and sodium content is 130 milligrams. Estimate the shelf probabilities for Apple Jacks.

```
## [1] 2
## Levels: 1 2 3 4
```

The model placed Apple Jacks on the 2nd shelf, which is the expected outcome!

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.



This plot shows that as the amount of sugar increases in the cereals selected, the probability of being on shelf two increases and the probability of being on shelves three or four decrease. This chart also shows that the amount of sugar in a cereal does not impact the probability of being on shelf one very much.

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.

```
##  sugar    fat sodium
##   0.27   0.30   0.23
```

```
## [1] "OR shelf 2 v 1"
```

```
##  sugar    fat sodium
##   2.06   3.37   0.02
```

```
##  sugar    fat sodium
##   0.48   0.30  55.74
```

```
## [1] "OR shelf 3 v 1"
```

```
##  sugar    fat sodium
##   0.04   0.85   0.00
```

```
##  sugar    fat sodium
##  26.81   1.18 311.36
```

```
## [1] "OR shelf 4 v 1"
```

```
##  sugar    fat sodium
##   0.05   0.77   0.00
```

```
##  sugar    fat sodium
##  21.48   1.30 290.31
```

```
##        lowr   upr
## sugar  0.14 29.68
## fat    0.87 13.04
## sodium 0.00  0.44
```

```
##        lowr  upr
## sugar  0.00 0.49
## fat    0.21 3.49
## sodium 0.00 0.12
```

```
##        lowr  upr
## sugar  0.00 0.61
## fat    0.19 3.16
## sodium 0.00 0.13
```

> Looking at the odds ratio values, we see trends that are similar to what we would expect given earlier analysis. The odds of a cereal being on shelf two increases 2.06 times if we increase sugar by one standard deviation, 0.27. With a 95% confidence, the odds of being on the second shelf rather than the first shelf change by 0.14 to 29.68 times when sugar is increased by 0.27, holding all other variables constant. If we decrease sugar by one standard deviation, the odds of a cereal being on the second shelf versus the first shelf is changed times 0.48. Reviewing the boxplots earlier in this report, we can see that cereals with higher sodium tend to be on the first shelf. The values comparing the odds of being on any shelf versus the first shelf show that if sodium is decreased by one standard deviation, a cereal has less odds of being on the first shelf and more odds of being on the second, third, or fourth shelves. The fat variable follows a similar trend to sugar. If the fat is increased by .30, the odds of being on the second shelf versus the first shelf change times 3.37 with a 95% confidence interval 0.87 to 13.04. After the second shelf, increases in fat lower the odds of a cereal being on the third or fourth 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:

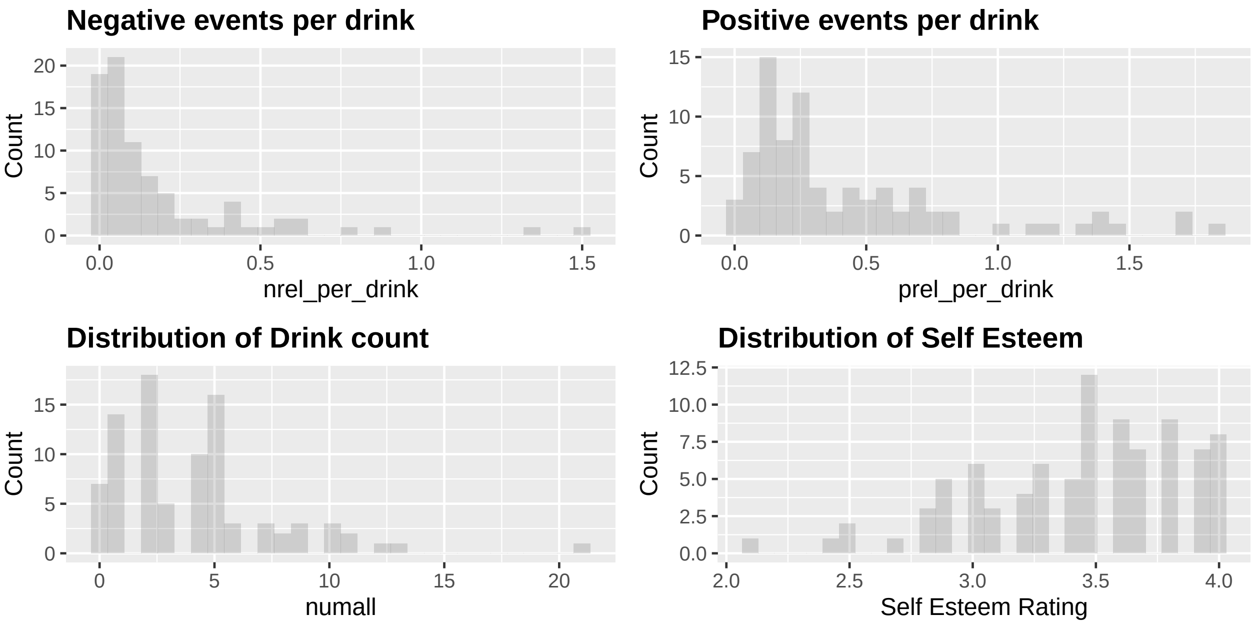
> *We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship*

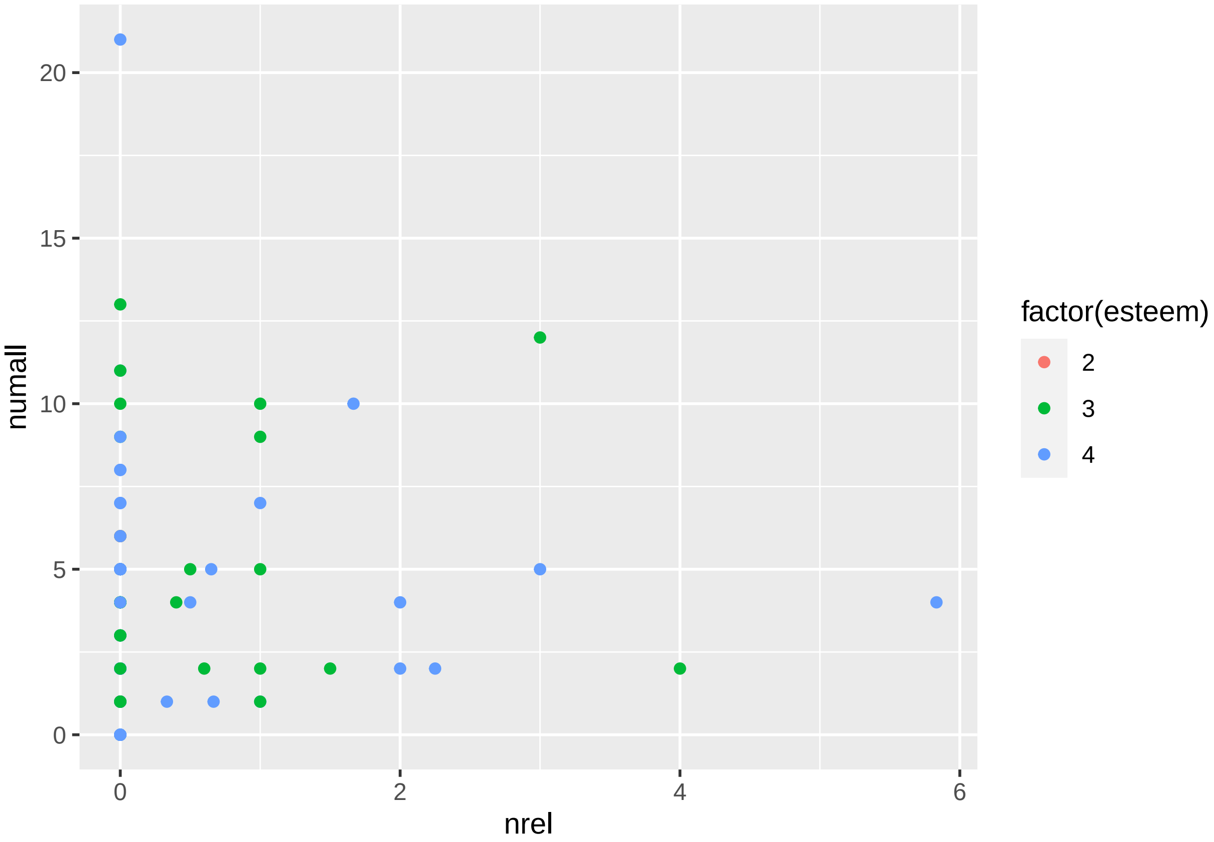
interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem.

```
## [1] 6 7 1 2 3 4 5
```

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.





>

Generally, people who do drink are most likely to participate on Saturdays. This is obviously not always the case, but most adults work between Monday through Friday and have Saturday and Sunday off. Because of this, our data is filtered to only include Saturday, with our goal being to better isolate the impact these variables have on the number of drinks consumed on a day most people have off. The distrobutions of negative and positive relationship events per drink look similar, but we can see that there are more instances of negative events under 0.5 drinks, but the number quickly drops off while positive events continue past the 0.5 mark. We also see that most people did not exceed five drinks on Saturday, and most ranked their self esteem between a 3-4. These plots don't demonstrate any obvious relationship between the count of drinks and the number of negative relationship events or self esteem rating.

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.

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = drinks2)
##
## 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
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  0.90934  1     0.3403
```

Using a log linked Poisson model we found that there is not a significant relationship between negative relationship events (`nrel`) and the number of drinks consumed. The coefficient for `nrel` in our model is 0.0497053. This coefficient translates to a 5.0961329 percent increase in the number of drinks per one standard deviation increase in negative relationship events, with a 95% confidence interval of -5.3888912, 15.4973085. This confidence interval contains zero, which aligns with `nrel` not showing a significant relationship with the number of drinks consumed as the impact of negative relationship events on drinking could lead to a negative or positive change.

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.

```
##
## Call:
## glm(formula = numall ~ rosn:nrel + nrel + rosn, family = poisson(link = "log"),
##     data = drinks2)
##
## 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
## rosn:nrel   -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
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##           LR Chisq Df Pr(>Chisq)
## nrel        1.0188  1    0.31281
## rosn        0.4122  1    0.52086
## rosn:nrel   4.7191  1    0.02983 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use a similar Poisson log-linked model to observe the relationships (if any) between self esteem `rosn`, negative relationship events `nrel`, and total number of drinks `numall`. In this model, we can see that both the `nrel` and `nrel:rosn` interaction terms have significance. To see which variable is impactiing this model more, we also ran an Anova test which showed that the interaction term is the more signififant of the two. This result rejects the null hypothesis that the interaction between self esteem and negative relationship events do not impact the total number of drinks consumed. We can also see that the coefficient for this interaction term is negative, which means the interaction term is inverseley related to `numall`. As self esteem or negative relationship events increase, the number of drinks is projected to decrease. More specifically, if `rosn:nrel` increases by one standard deviation, the number of drinks is changed by -24.97%.