W271 Spring 18: Lab 2

Alyssa Eisenberg, Jeffrey Hsu, Gerard Kelly

Alcohol Consumption, Self-Esteem and Romantic Interactions

Introduction

The researchers stated the hypothesis as follow: *"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."*

Gerard EDA

```
library(car); require(dplyr); library(Hmisc); library(mcprofile); library(ggplot2); library(gridExtra);
dehart <- read.table(file="DeHartSimplified.csv", header=TRUE, sep=",")
#describe(dehart) #with a 10-page limit, should we include this type of output?
```

The dataset contains 623 observations of 13 variables representing entries in records kept by study participants. The variable **id** is a numeric identifier for each of the 89 study participants. Each participant recorded entries for seven consecutive days, indexed by the **studyday** variable, with the **dayweek** variable indicating which days of the week these correspond to (Monday = 1). The variable **gender** takes on one of two values according to whether the participant is male (1) or female (2); about 56% of the participants are female.

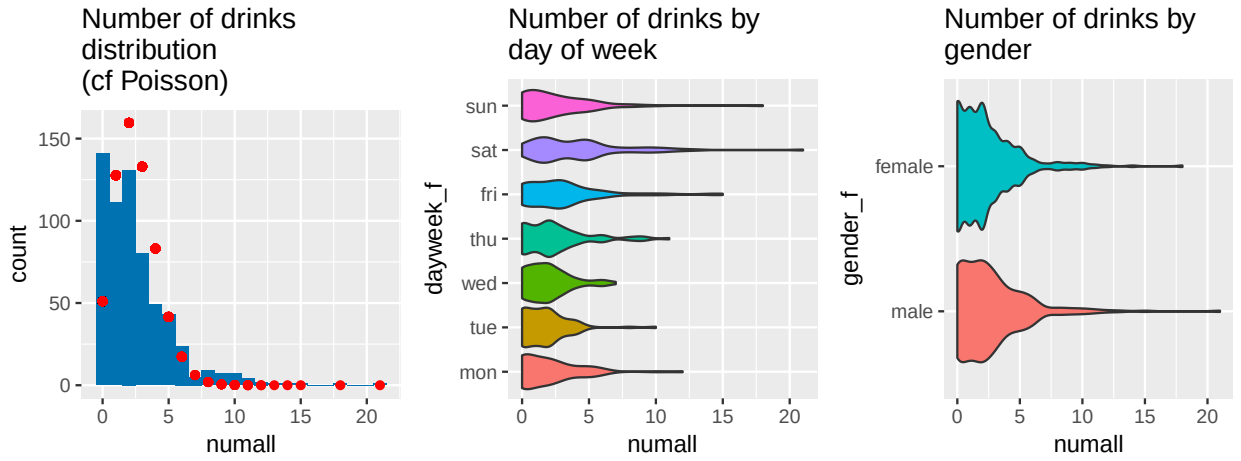
```
dehart$dayweek_f <- factor(dehart$dayweek); levels(dehart$dayweek_f) = c("mon","tue","wed","thu","fri",
dehart$gender_f <- factor(dehart$gender); levels(dehart$gender_f) = c("male","female")
```

For each of the seven days, participants record the number of drinks consumed with the integer **numall** count variable. There is one missing value. Values range from 0 to 15 with single outliers at 19 and 21. Observations are concentrated in the range 0 to 5. The sample mean and sample variance are 2.52 and 2.66 respectively. The median number of drinks is 3 for Friday, 4 for Saturday and 2 for all other days. There is a pronounced positive skew to the number of drinks for Fridays, Saturdays and Sundays.

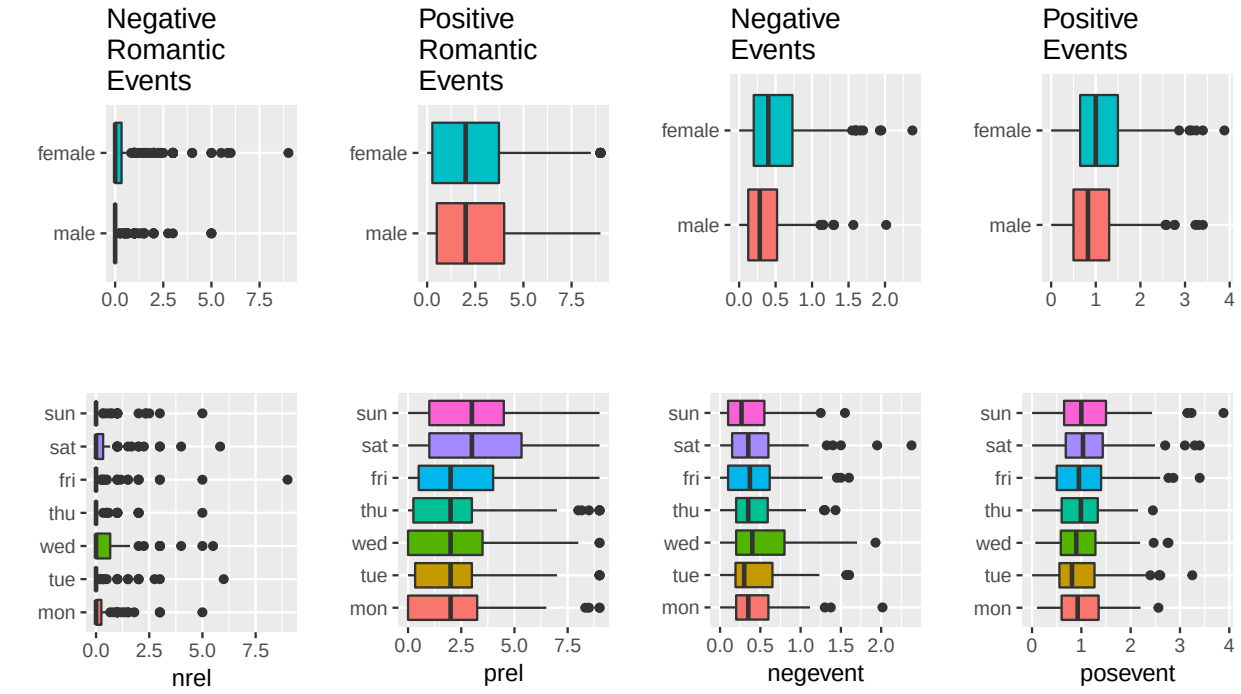
The variables **nrel** and **prel** are index measures for the negative and positive romantic-relationship events experienced by the subject each day (ranging from 0 to around 9), while **negevent** and **posevent** are index values combining the total number and intensity of negative and positive events experienced that day (ranging from 0 to around 4). There are no missing values for these variables. Distributions (and scales) for these measurements are very different, although tend to be similar for males and females, and similar across days of the week, although **prel** has signifiantly higher average (mean and median) values on Saturdays and Sundays. Distributions for all these variables are strongly positively-skewed, but this is particularly extreme for **nrel**, where a small fraction of outlying observations accounts for almost all of the variation.

```
#mean(dehart$numall, na.rm=TRUE); sd(dehart$numall, na.rm=TRUE)
#aggregate(numall ~ dayweek_f, data = dehart, FUN = function(x) c(m = mean(x), n = median(x)))
#The distribution of 622 times a Poisson random variable with parameter lambda is overlaid on the histo
lambda = 2.5
p1 <- ggplot(na.omit(dehart), aes(x = numall)) + geom_histogram(aes(y = ..count..), binwidth = 1, fill=
  geom_point(aes(y = 622*dpois(x = numall,lambda)), color = "red")+ ggtitle("Number of drinks\ndistribu
```

```
p2<-ggplot(na.omit(dehart), aes(dayweek_f, numall)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Numb
p3<-ggplot(na.omit(dehart), aes(gender_f, numall)) + geom_violin(aes(fill = gender_f)) + ggtitle("Numbe
grid.arrange(p1, p2, p3, ncol = 3)
```




```
p1a<-ggplot(dehart, aes(gender_f, nrel)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = "") + g
p1b<-ggplot(dehart, aes(dayweek_f, nrel)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + ggtitle
p2a<-ggplot(dehart, aes(gender_f, prel)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = "") + g
p2b<-ggplot(dehart, aes(dayweek_f, prel)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + ggtitle
p3a<-ggplot(dehart, aes(gender_f, negevent)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = ""
p3b<-ggplot(dehart, aes(dayweek_f, negevent)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + gg
p4a<-ggplot(dehart, aes(gender_f, posevent)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = ""
p4b<-ggplot(dehart, aes(dayweek_f, posevent)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + gg
grid.arrange(p1a, p2a, p3a, p4a, p1b, p2b, p3b, p4b, ncol = 4)
```




```
aggregate(nrel ~ dayweek_f, data = dehart, FUN = mean)
```

```
##   dayweek_f      nrel
## 1       mon 0.3471910
## 2       tue 0.2882022
## 3       wed 0.5541466
## 4       thu 0.2295880
## 5       fri 0.4014981
## 6       sat 0.4033708
## 7       sun 0.2893258
```

The **rosn** variable measures trait (long-term) self-esteem, a single measurement for each participant taken at the beginning of the study that does not change over the course of the seven days. This measurement ranges between 2 and 4, with a mean value around 3.4. Distributions for males and females differ, with a larger proportion of males recording lower values and a larger proportion of females recording higher values. The **age** variable measures age in years, ranging between 24.4 and 42.3 with a mean value of 34.3 and similar distributions for males and females but with a higher proportion of females recording higher values. Neither **rosn** nor **age** have any missing observations.

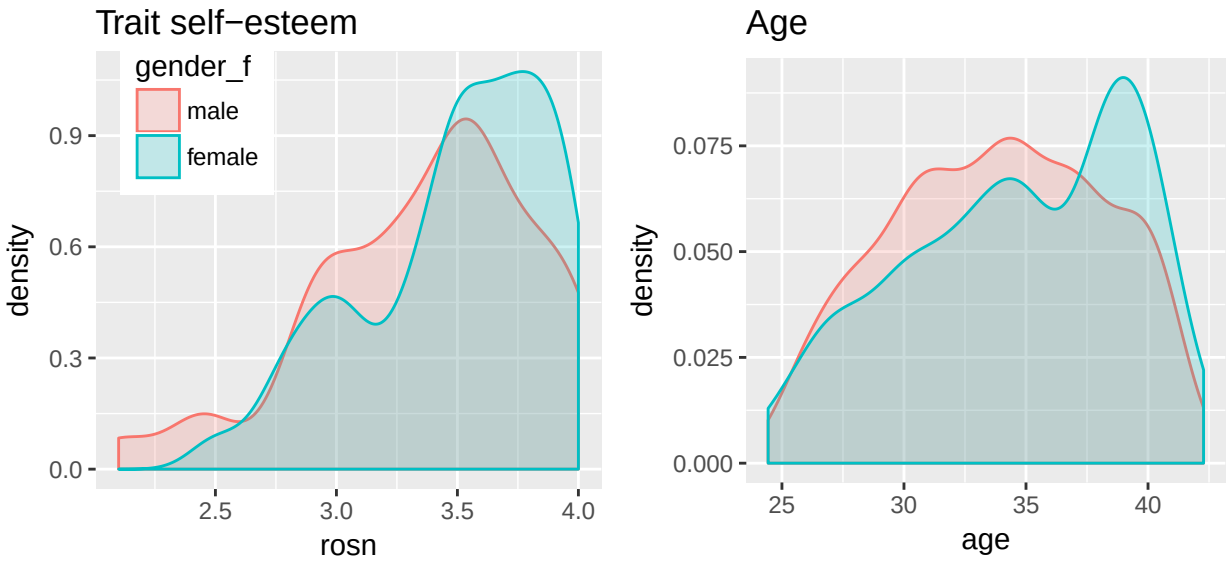
```
quantile(dehart$rosn[dehart$gender == 1])
```

```
##   0%  25%  50%  75% 100%
##  2.1  3.0  3.5  3.7  4.0
```

```
quantile(dehart$rosn[dehart$gender == 2])
```

```
##   0%  25%  50%  75% 100%
## 2.50 3.30 3.55 3.80 4.00
```

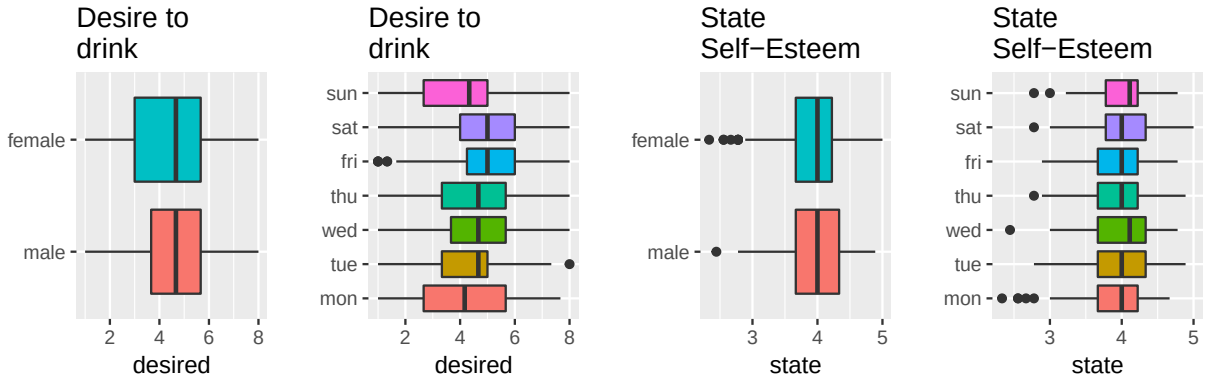
```
p1<-ggplot(dehart, aes(x = rosn, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+ ggtitl
p2<-ggplot(dehart, aes(x = age, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+
  ggtitle("Age")+theme(legend.position="none")
grid.arrange(p1, p2, ncol = 2)
```



The **desired** variable is a measure of the participant's recorded desire to drink, with values ranging between 1 and 8, a mean of 4.5 and a fairly symmetric distribution. with a significant share of responses at minimum and maximum values. Average values are slightly higher on average for males than for females and are highest on Friday and Saturday and lowest on Sunday and Monday. The **state** variable is a record of the participant's state (short-term) self-esteem as it varies each day. This ranges between 2 and 5 with a mean of 3.97 and a moderately negative-skewed distribution, little difference between males and females and little

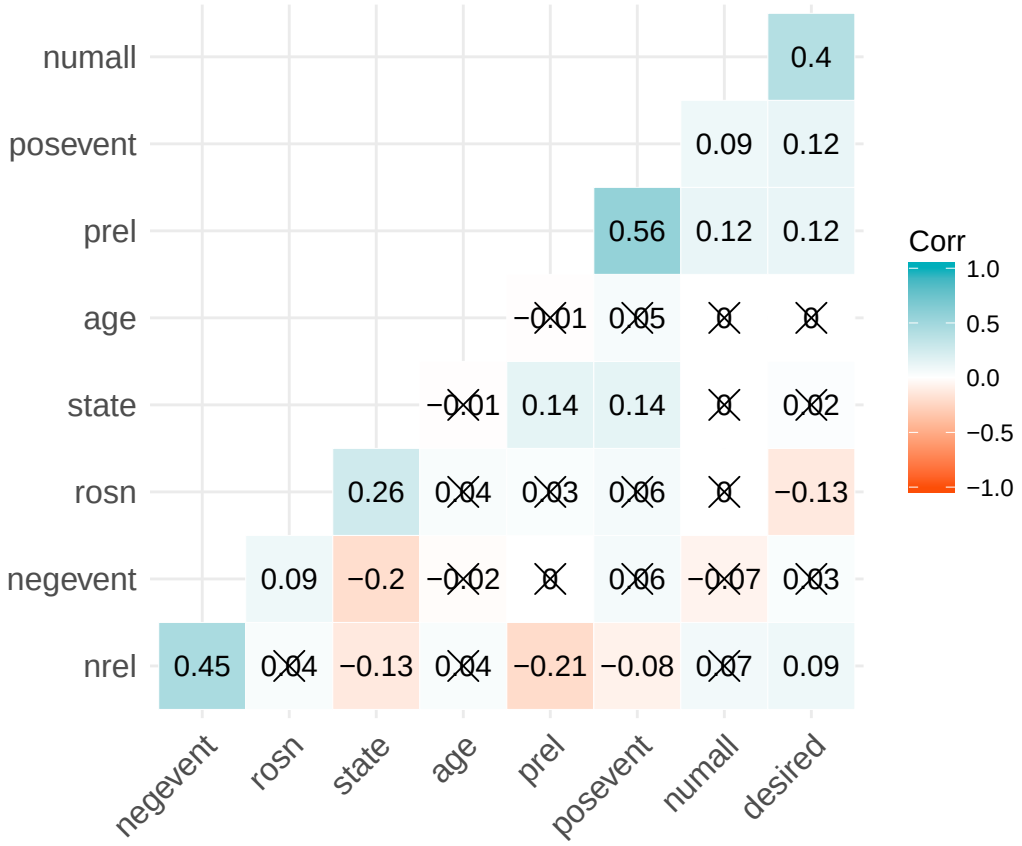
difference between days of the week (although with more negative outliers on Mondays). The variables **desired** and **state** both contain three missing values, two of these being from the same observation.

```
# aggregate(desired ~ gender_f, data = dehart, FUN = mean)
# aggregate(desired ~ dayweek_f, data = dehart, FUN = mean)
# aggregate(state ~ gender_f, data = dehart, FUN = mean)
# aggregate(state ~ dayweek_f, data = dehart, FUN = mean)
p1<-ggplot(na.omit(dehart), aes(gender_f, desired)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "")
  ggtitle("Desire to\ndrink") + theme(legend.position="none") + coord_flip()
p2<-ggplot(na.omit(dehart), aes(dayweek_f, desired)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "
  ggtitle("Desire to\ndrink") + theme(legend.position="none") + coord_flip()
p3<-ggplot(na.omit(dehart), aes(gender_f, state)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "") +
  ggtitle("State\nSelf-Esteem") + theme(legend.position="none") + coord_flip()
p4<-ggplot(na.omit(dehart), aes(dayweek_f, state)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "")
  ggtitle("State\nSelf-Esteem") + theme(legend.position="none") + coord_flip()
grid.arrange(p1, p2, p3, p4, ncol = 4)
```



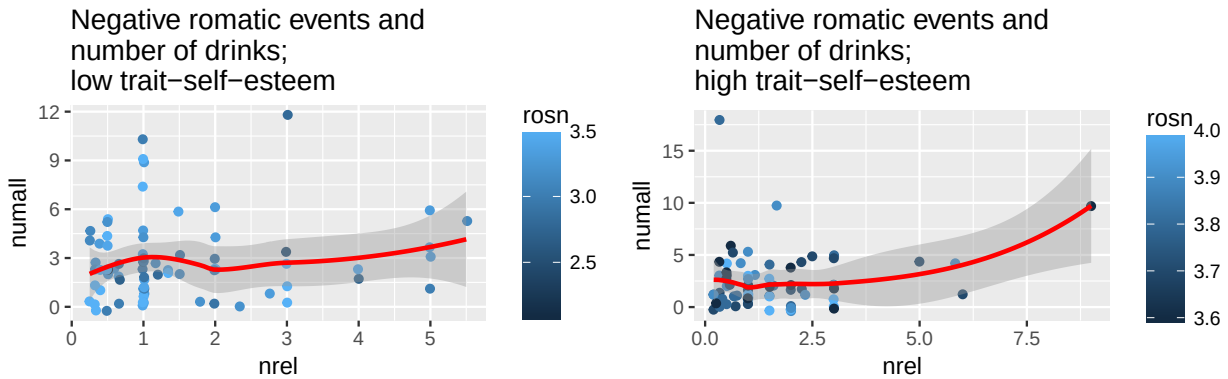
Bivarate relationships between these continuous variables are summarized in the below plot of correlation coefficients below. Most of these relationship are weak or insignificant. The strongest positive correlations are between **nrel** and **negevent** (negative events and negative romantic events), and between **prel** and **posevent** (positive events and positive romantic events). There is also a moderate positive correlation between the number of drinks and the desire to drink (**numall** and **desired**), and a mild positive correlation between trait and state self-esteem (**rosn** and **state**). State self-esteem has a weak positive correlation to positive event variables and a weak negative correlation to negative event variables, but no correlation with the number of drinks nor the desire to drink. Trait self-esteem has a weak negative correlation with the desire to drink but no correlation with the number of drinks. Age is uncorreated with any other variable.

```
data <- na.omit(dehart[,c(4,5,6,7,8,10,11,12,13)])
corr <- round(cor(data), 2)
ggcorrplot(corr, p.mat = cor_pmat(data),hc.order = TRUE, type = "lower", color = c("#FC4E07", "white",
```



The heavily skewed distribution for **nrel** indicates that the negative romatic relationship events relevant to the hypothesis are relatively infrequent. A subset of the dataset can be created for participant-days involving a non-zero rating on this variable. The relationship between **nrel** and numall can then be compared for participants with below-median trait self-esteem and above-median trait self-esteem. The relationship between negative romatic relationship events and the number of drinks appears to be stronger for individuals with higher trait-self esteem, however the range of this is driven by a relatively small number of individuals recording high **nrel** values that do not appear in the lower trait-self-esteem subset.

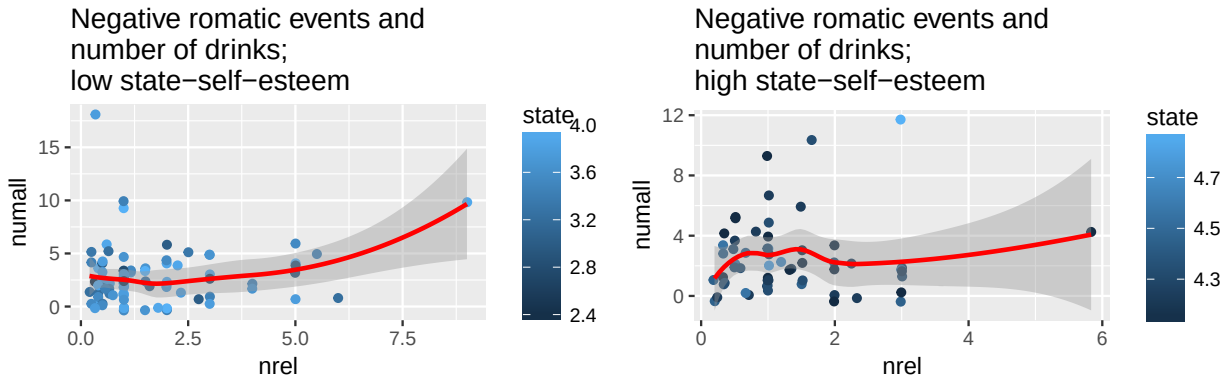
```
dehart_nrel = dehart[which(dehart$nrel != 0),]
p1 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$rosn<=3.5),]), aes(nrel, numall)) + geom_jitter(aes(
p2 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$rosn>3.5),]), aes(nrel, numall)) + geom_jitter(aes(c
grid.arrange(p1, p2,ncol = 2)
```



The same comparison can be made using state self-esteem rather than trait self-esteem, however since state-self-esteem is affected by negtive romantic relationship events, this comparson may be less pertinent to

the hypothesis.

```
p1 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$state<=4),]), aes(nrel, numall)) + geom_jitter(aes(c
p2 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$state>4),]), aes(nrel, numall)) + geom_jitter(aes(col
grid.arrange(p1, p2,ncol = 2)
```



Alyssa EDA

Note that I am using the setup from Gerard's EDA to keep consistent (reading in the data file, libraries used).

Let us first examine the summary of the dataset and understand our variables.

```
# Overview of data
describe(dehart)
```

```
## dehart
##
##  15  Variables      623  Observations
## ---------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       89        1    75.89    56.82      7.0     16.2
##      .25      .50      .75      .90      .95
##     33.0     60.0    123.0    147.2    153.0
##
## lowest :   1   2   4   5   7, highest: 153 154 155 156 160
## ---------------------------------------------------------------------------
## studyday
##        n  missing distinct     Info     Mean      Gmd
##      623        0        7     0.98        4    2.289
##
## Value          1     2     3     4     5     6     7
## Frequency     89    89    89    89    89    89    89
## Proportion 0.143 0.143 0.143 0.143 0.143 0.143 0.143
## ---------------------------------------------------------------------------
## dayweek
##        n  missing distinct     Info     Mean      Gmd
##      623        0        7     0.98        4    2.289
##
## Value          1     2     3     4     5     6     7
## Frequency     89    89    89    89    89    89    89
## Proportion 0.143 0.143 0.143 0.143 0.143 0.143 0.143
```

```
## ------------------------------------------------------------------------
## numall
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      622        1       18     0.97    2.524    2.636     0.00     0.00
##      .25      .50      .75      .90      .95
##     1.00     2.00     3.75     6.00     8.00
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency    141   112   132    81    49    43    24     6     9     7
## Proportion 0.227 0.180 0.212 0.130 0.079 0.069 0.039 0.010 0.014 0.011
##
## Value         10    11    12    13    14    15    18    21
## Frequency      7     4     2     1     1     1     1     1
## Proportion 0.011 0.006 0.003 0.002 0.002 0.002 0.002 0.002
## ------------------------------------------------------------------------
## nrel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       33    0.551    0.359   0.6252        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        1        2
##
## lowest : 0.0000000 0.2000000 0.2500000 0.3333333 0.4000000
## highest: 5.0000000 5.5000000 5.8333333 6.0000000 9.0000000
## ------------------------------------------------------------------------
## prel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       68    0.982    2.583    2.613   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.4167   2.0000   4.0000   6.0000   7.8683
##
## lowest : 0.0000000 0.2000000 0.2500000 0.3333333 0.5000000
## highest: 8.1666667 8.3333333 8.5000000 8.6666667 9.0000000
## ------------------------------------------------------------------------
## negevent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0      131    0.996   0.4414   0.4123   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.1583   0.3500   0.6292   1.0000   1.1500
##
## lowest : 0.00000000 0.02500000 0.03333333 0.05000000 0.07500000
## highest: 1.70000000 1.93000000 1.95000000 2.01666667 2.37666667
## ------------------------------------------------------------------------
## posevent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0      216        1    1.048   0.7077    0.200    0.300
##      .25      .50      .75      .90      .95
##    0.600    0.950    1.378    1.938    2.200
##
## lowest : 0.00000000 0.04000000 0.05000000 0.06666667 0.10000000
## highest: 3.23333333 3.25000000 3.30000000 3.40000000 3.88333333
## ------------------------------------------------------------------------
## gender
##        n  missing distinct     Info     Mean      Gmd
##      623        0        2    0.739    1.562   0.4932
```

```
##
## Value          1     2
## Frequency    273   350
## Proportion 0.438 0.562
## ---------------------------------------------------------------------------
## rosn
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       17    0.993    3.436   0.4663      2.7      2.9
##      .25      .50      .75      .90      .95
##      3.2      3.5      3.8      3.9      4.0
##
## Value        2.1   2.4   2.5   2.7   2.8   2.9   3.0   3.1   3.2   3.3
## Frequency      7     7    14     7    21    35    42    21    28    42
## Proportion 0.011 0.011 0.022 0.011 0.034 0.056 0.067 0.034 0.045 0.067
##
## Value        3.4   3.5   3.6   3.7   3.8   3.9   4.0
## Frequency     35    84    63    49    63    49    56
## Proportion 0.056 0.135 0.101 0.079 0.101 0.079 0.090
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       89        1    34.29     5.18    26.24    27.82
##      .25      .50      .75      .90      .95
##    30.53    34.57    38.19    40.15    40.56
##
## lowest : 24.43258 25.57700 26.05613 26.14100 26.23682
## highest: 40.56400 40.58864 40.68720 40.82957 42.27789
## ---------------------------------------------------------------------------
## desired
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      620        3       22    0.996    4.465    1.921    1.333    2.000
##      .25      .50      .75      .90      .95
##    3.333    4.667    5.667    6.667    7.333
##
## lowest : 1.000000 1.333333 1.666667 2.000000 2.333333
## highest: 6.666667 7.000000 7.333333 7.666667 8.000000
## ---------------------------------------------------------------------------
## state
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      620        3       25    0.993    3.966   0.4894    3.222    3.333
##      .25      .50      .75      .90      .95
##    3.667    4.000    4.222    4.556    4.556
##
## lowest : 2.333333 2.444444 2.555556 2.666667 2.777778
## highest: 4.555556 4.666667 4.777778 4.888889 5.000000
## ---------------------------------------------------------------------------
## dayweek_f
##        n  missing distinct
##      623        0        7
##
## Value        mon   tue   wed   thu   fri   sat   sun
## Frequency     89    89    89    89    89    89    89
## Proportion 0.143 0.143 0.143 0.143 0.143 0.143 0.143
## ---------------------------------------------------------------------------
```

```
## gender_f
##        n  missing distinct
##      623        0        2
##
## Value        male female
## Frequency     273    350
## Proportion  0.438  0.562
## ---------------------------------------------------------------------------
```

```
# Gather more data on count of data points per participant, and how their study day vs week day relate
table((table(dehart$id)))
```

```
##
##  7
## 89
```

```
xtabs(~studyday + dayweek, data=dehart)
```

```
##         dayweek
## studyday  1  2  3  4  5  6  7
##        1 10  7 19 15 16  6 16
##        2 16 10  7 19 15 16  6
##        3  6 16 10  7 19 15 16
##        4 16  6 16 10  7 19 15
##        5 15 16  6 16 10  7 19
##        6 19 15 16  6 16 10  7
##        7  7 19 15 16  6 16 10
```

```
# Examine the participants with any missing values
dehart$id[is.na(dehart$numall)]
```

```
## [1] 42
```

```
dehart$id[is.na(dehart$desired)]
```

```
## [1]   2 110 116
```

```
dehart$id[is.na(dehart$state)]
```

```
## [1]   2   4 110
```

```
dehart[dehart$id %in% dehart$id[is.na(dehart$numall)],]
```

```
##     id studyday dayweek numall nrel prel negevent posevent gender rosn
## 211 42        1       4      6    0    6     0.00     1.30      2    4
## 212 42        2       5      4    0    5     0.50     1.60      2    4
## 213 42        3       6      3    0    6     0.90     2.10      2    4
## 214 42        4       7     NA    0    3     0.00     1.80      2    4
## 215 42        5       1      5    0    3     0.15     1.35      2    4
## 216 42        6       2      0    0    3     0.80     0.60      2    4
## 217 42        7       3      3    0    3     0.60     0.90      2    4
##          age  desired    state dayweek_f gender_f
## 211 35.15674 4.666667 4.333333       thu   female
## 212 35.15674 4.666667 4.444444       fri   female
## 213 35.15674 3.666667 4.555556       sat   female
## 214 35.15674 3.666667 4.555556       sun   female
## 215 35.15674 3.333333 4.555556       mon   female
## 216 35.15674 1.000000 4.555556       tue   female
## 217 35.15674 3.666667 4.555556       wed   female
```

```
dehart[dehart$id %in% dehart$id[is.na(dehart$desired) | is.na(dehart$state)],]
```

```
##      id studyday dayweek numall      nrel      prel  negevent  posevent
## 8     2        1       3      3 1.1666667 4.5000000 1.6566667 2.1366667
## 9     2        2       4      4 0.0000000 5.5000000 1.0666667 2.0616667
## 10    2        3       5      0 2.0000000 3.0000000 0.6678571 1.9385714
## 11    2        4       6      4 5.8333333 0.8333333 2.3766667 0.9241667
## 12    2        5       7      7 0.0000000 0.0000000 0.0000000 0.0000000
## 13    2        6       1      4 0.8333333 6.5000000 0.8250000 1.6416667
## 14    2        7       2      1 0.0000000 9.0000000 1.1300000 2.3983333
## 15    4        1       3      0 0.7142857 3.8571429 0.5214286 1.0107143
## 16    4        2       4      1 0.0000000 2.0000000 0.3000000 1.4083333
## 17    4        3       5      3 0.2500000 6.0000000 0.5716667 1.4200000
## 18    4        4       6      1 0.3333333 4.0000000 0.2333333 1.3464286
## 19    4        5       7      0 0.0000000 5.6666667 0.2583333 1.9916667
## 20    4        6       1      1 0.3333333 1.0000000 0.4333333 2.0500000
## 21    4        7       2      0 0.2000000 1.8000000 0.5200000 0.8800000
## 400 110        1       6      2 2.0000000 0.0000000 0.4500000 0.2000000
## 401 110        2       7      2 0.0000000 0.0000000 0.0000000 0.4000000
## 402 110        3       1      1 0.0000000 0.0000000 0.1000000 0.7000000
## 403 110        4       2      1 0.0000000 0.0000000 0.2000000 0.6000000
## 404 110        5       3      2 3.0000000 0.0000000 0.3000000 0.8000000
## 405 110        6       4      2 0.0000000 0.0000000 0.2000000 0.2000000
## 406 110        7       5      2 0.0000000 0.0000000 0.0000000 0.4000000
## 442 116        1       4      3 0.0000000 2.0000000 0.4000000 1.4500000
## 443 116        2       5      2 0.0000000 2.0000000 0.0000000 0.9000000
## 444 116        3       6      3 0.0000000 5.0000000 0.1000000 2.0000000
## 445 116        4       7      2 0.0000000 4.0000000 0.3000000 1.2000000
## 446 116        5       1      1 0.0000000 2.0000000 0.0000000 1.5500000
## 447 116        6       2      2 0.0000000 2.0000000 0.1500000 1.3000000
## 448 116        7       3      2 0.0000000 2.0000000 0.2000000 1.3000000
##     gender rosn      age  desired    state dayweek_f gender_f
## 8        2  3.9 38.00137 3.666667 3.666667       wed   female
## 9        2  3.9 38.00137 4.666667 4.111111       thu   female
## 10       2  3.9 38.00137 5.000000 3.666667       fri   female
## 11       2  3.9 38.00137 5.666667 4.111111       sat   female
## 12       2  3.9 38.00137       NA       NA       sun   female
## 13       2  3.9 38.00137 4.333333 4.222222       mon   female
## 14       2  3.9 38.00137 2.666667 3.888889       tue   female
## 15       2  3.7 30.04791 1.666667 4.222222       wed   female
## 16       2  3.7 30.04791 7.666667 4.000000       thu   female
## 17       2  3.7 30.04791 5.666667       NA       fri   female
## 18       2  3.7 30.04791 5.000000 4.111111       sat   female
## 19       2  3.7 30.04791 4.666667 4.222222       sun   female
## 20       2  3.7 30.04791 5.000000 4.333333       mon   female
## 21       2  3.7 30.04791 4.666667 4.444444       tue   female
## 400      2  3.6 40.82957 3.333333 4.111111       sat   female
## 401      2  3.6 40.82957 4.666667 4.222222       sun   female
## 402      2  3.6 40.82957       NA       NA       mon   female
## 403      2  3.6 40.82957 1.333333 4.333333       tue   female
## 404      2  3.6 40.82957 5.000000 4.222222       wed   female
## 405      2  3.6 40.82957 4.333333 4.222222       thu   female
## 406      2  3.6 40.82957 3.333333 4.444444       fri   female
## 442      2  3.4 37.38809 7.333333 3.666667       thu   female
```

```
## 443      2  3.4 37.38809 5.000000 4.000000       fri   female
## 444      2  3.4 37.38809 6.000000 3.888889       sat   female
## 445      2  3.4 37.38809 5.000000 4.111111       sun   female
## 446      2  3.4 37.38809 4.666667 4.000000       mon   female
## 447      2  3.4 37.38809 5.000000 4.111111       tue   female
## 448      2  3.4 37.38809       NA 4.000000       wed   female
```

```
# Check on ptoentially erroneous values
dehart[dehart$posevent>3,]
```

```
##      id studyday dayweek numall nrel     prel  negevent posevent gender
## 44   10        2       6      2    0 7.000000 0.0000000 3.400000      2
## 424 113        4       5      9    0 5.500000 0.6500000 3.400000      1
## 425 113        5       6     12    3 6.000000 0.4000000 3.300000      1
## 426 113        6       7      5    0 8.333333 0.0000000 3.233333      1
## 452 118        4       7      8    0 9.000000 0.3000000 3.150000      2
## 540 140        1       7      2    0 0.000000 0.2333333 3.883333      2
## 542 140        3       2      1    0 3.000000 0.1866667 3.250000      2
## 546 140        7       6      4    0 6.000000 0.6500000 3.100000      2
##     rosn      age  desired    state dayweek_f gender_f
## 44   3.5 38.19028 4.666667 4.777778       sat   female
## 424  2.8 28.00548 6.000000 4.777778       fri     male
## 425  2.8 28.00548 7.000000 4.888889       sat     male
## 426  2.8 28.00548 6.333333 4.777778       sun     male
## 452  2.8 32.83231 5.666667 3.777778       sun   female
## 540  3.9 38.96783 1.666667 4.000000       sun   female
## 542  3.9 38.96783 4.333333 3.888889       tue   female
## 546  3.9 38.96783 6.000000 4.222222       sat   female
```

This dataset has 623 observations and 13 variables. There are a few missing values and potential erroneous values, noted at each variable below where they occur. There are no values that appear top or bottom coded in the data.

Variables

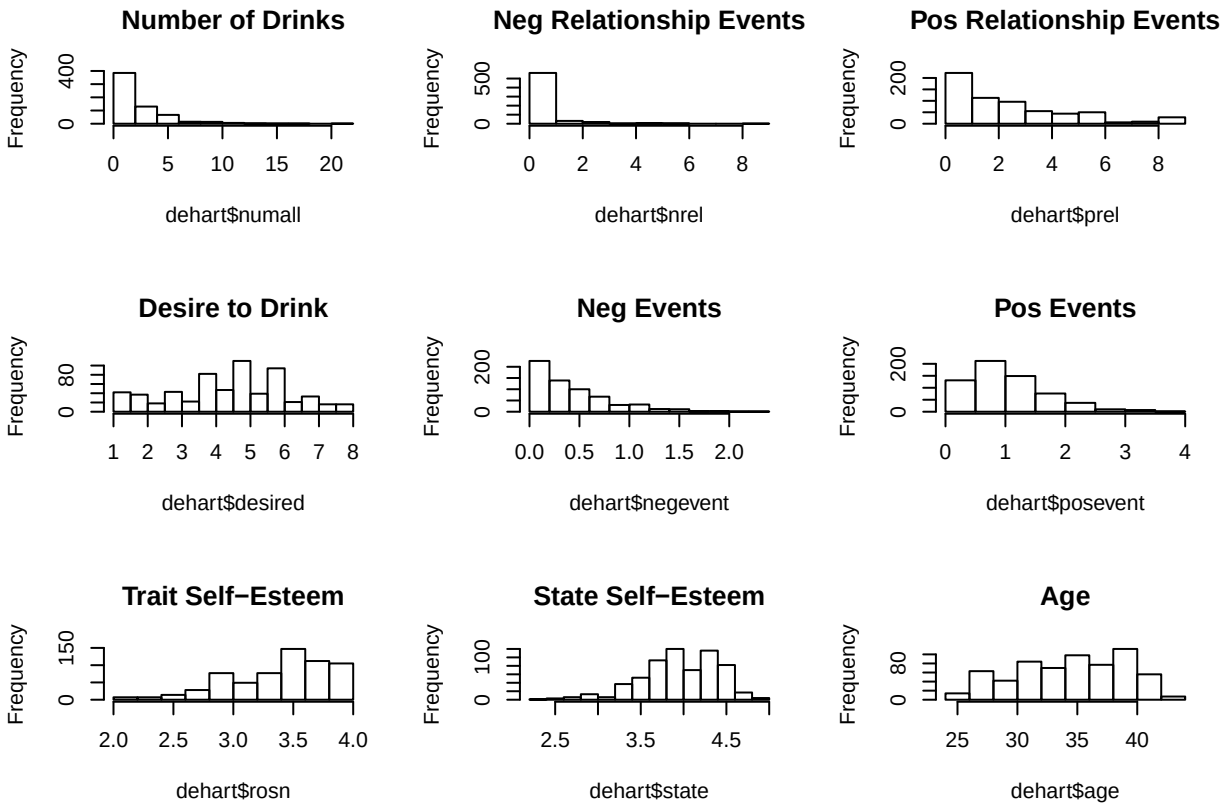
- Each unique participant is assigned an id number, encoded as **id**. There are 89 unique participants in the dataset, each with 7 data points
- *studyday* encodes which day of the study it was for the participant. There are 89 observations for each of study days 1 through 7, meaning that we have data for each participant for their first seven days of the study
- *dayweek* is the day of the week for the observation. Again, there are 89 observations for each day of the week 1 through 7 (Monday is coded as 1). Note that participants are unevenly spaced out for which day of the week they start the study: 10 start on Monday, 7 on Tuesday, 19 on Wednesday, 15 on Thursday, 16 on Friday, 6 on Saturday, and 16 on Sunday
- *numall* is the number of alcoholic drinks consumed on that day. It is an integer variable taking 18 distinct values, ranging from 0 to 21 drinks. Note that there is 1 missing value for participant id 42.
- *nrel* is a measure of negative romantic relationship interactions. It is a continuous variable taking 33 distinct values, ranging from 0 to 9
- *prel* is a measure of positive romantic relationship interactions. It is a continuous variable taking 68 distinct values, ranging from 0 to 9
- *negevent* is a combination of several items scored on a 0-3 scale measuring the total number and intensity of negative events on the given day. A higher value indicates a larger number of negative events and/or

more extremely negative events. It is a continuous variable taking 131 distinct values, ranging from 0 to 2.4

- *posevent* is a combination of several items scored on a 0-3 scale measuring the total number and intensity of positive events on the given day. A higher value indicates a larger number of positive events and/or more extremely positive events. It is a continuous variable taking 216 distinct values, ranging from 0 to 3.9. Values above 3 are a bit suspicious since this variable combines several items scored on a 0-3 scale, but we cannot be sure that these are erroneous values since we do not know the combination formula or the individual values. There are only 8 rows with values over 3, and the rest of the data appears normal. Thus, we will continue our investigation assuming these are valid data points
- *gender* is coded as 1 (male) or 2 (female), with a slightly higher proportion of females (39 males, and 50 females)
- *rosn* is our measure for trait self-esteem, which is a long-term view of self-worth. This value was measured once at the beginning of the study, so the same value carries through all seven observations for each individual. It is a continuous variable taking 17 distinct values ranging from 2.1 to 4
- *age* is a continuous variable taking 89 distinct values ranging from 24.4 to 42.3
- *desired* measures the participants' desire to drink, with a higher score meaning a greater desire. It is a continuous variable taking 22 distinct values ranging from 1 to 8. Note that there are 3 missing values for participant ids 2, 110, 116.
- *state* is our measure for state self-esteem, which is a short-term view of self-worth. This was measured daily, unlike *rosn* long-term self-esteem. It is a continuous variable taking 25 distinct values ranging from 2.3 to 5. Note that there are 3 missing values for participant ids 2, 4, 110.

Now that we understand the basic about our data, let us examine the univariate distributions. We already know the distributions for participant id, study day, day of the week, and gender. Thus, we will focus here on our nine other continuous variables.

```
# Distributions - histograms
par(mfrow=c(3,3))
hist(dehart$numall, main = "Number of Drinks")
hist(dehart$nrel, main = "Neg Relationship Events")
hist(dehart$prel, main = "Pos Relationship Events")
hist(dehart$desired, main = "Desire to Drink")
hist(dehart$negevent, main = "Neg Events")
hist(dehart$posevent, main = "Pos Events")
hist(dehart$rosn, main = "Trait Self-Esteem")
hist(dehart$state, main = "State Self-Esteem")
hist(dehart$age, main = "Age")
```




```
# Distributions - table format for more detail on outcome variables
xtabs(~numall, data=dehart) #Number of drinks
```

```
## numall
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  18  21
## 141 112 132  81  49  43  24   6   9   7   7   4   2   1   1   1   1   1
```

```
round(prop.table(xtabs(~numall, data=dehart)), 2) #Proportion of number of drinks
```

```
## numall
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
## 0.23 0.18 0.21 0.13 0.08 0.07 0.04 0.01 0.01 0.01 0.01 0.01 0.00 0.00 0.00
##   15   18   21
## 0.00 0.00 0.00
```

```
xtabs(~desired, data=dehart) #Desire to drink
```

```
## desired
##           1 1.333333333 1.666666667           2 2.333333333 2.666666667
##          26          16          11          26          18          20
##           3 3.333333333 3.666666667           4 4.333333333 4.666666667
##          23          22          34          48          47          49
##           5 5.333333333 5.666666667           6 6.333333333 6.666666667
##          61          39          45          49          21          17
##           7 7.333333333 7.666666667           8
##          16          16           6          10
```

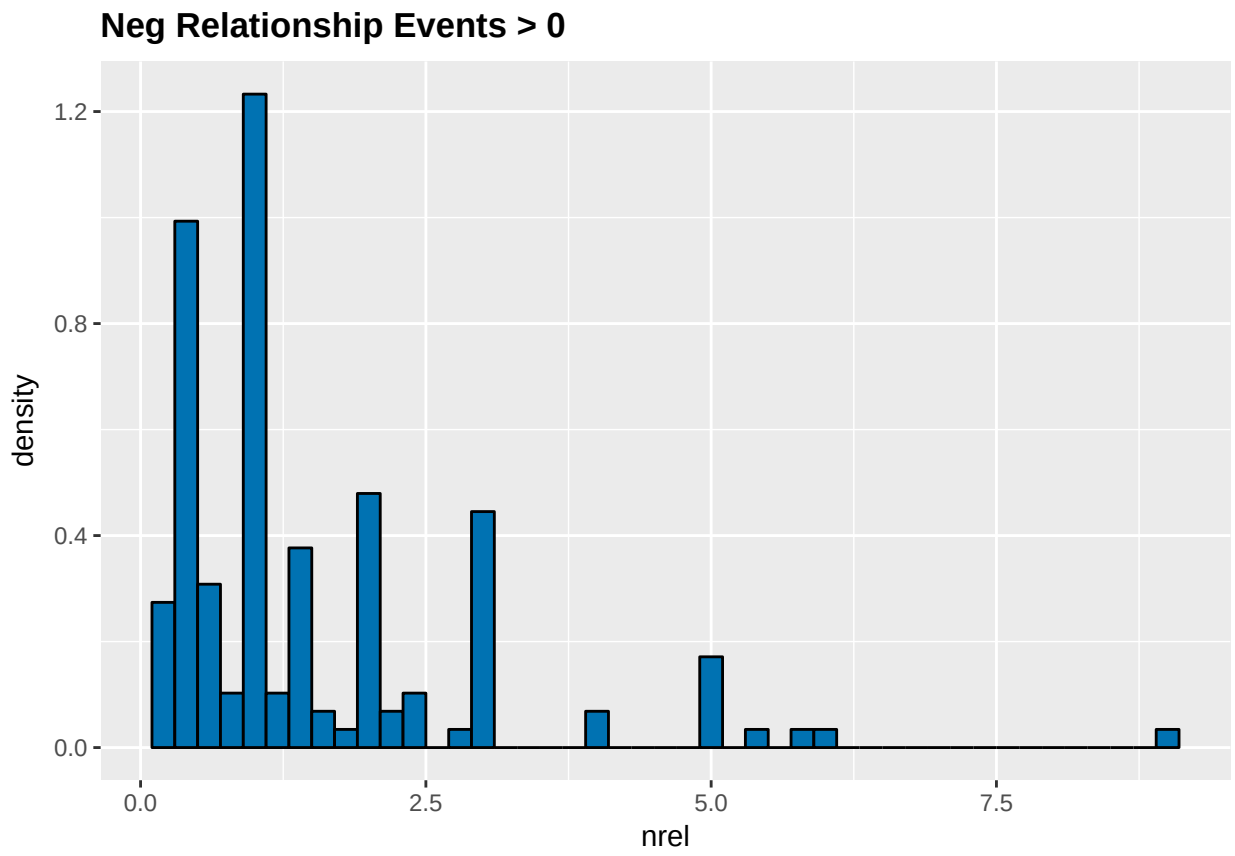
```
round(prop.table(xtabs(~desired, data=dehart)), 2) #Proportion of desire to drink
```

```
## desired
##           1 1.333333333 1.666666667           2 2.333333333 2.666666667
##        0.04        0.03        0.02        0.04        0.03        0.03
##           3 3.333333333 3.666666667           4 4.333333333 4.666666667
##        0.04        0.04        0.05        0.08        0.08        0.08
##           5 5.333333333 5.666666667           6 6.333333333 6.666666667
##        0.10        0.06        0.07        0.08        0.03        0.03
##           7 7.333333333 7.666666667           8
##        0.03        0.03        0.01        0.02
```

```
# Deeper dive on distribution of particular interest: negative relationship events
round(prop.table(xtabs(~nrel, data=dehart))[1], 2)
```

```
##    0
## 0.77
```

```
ggplot(dehart[dehart$nrel>0,], aes(x = nrel)) +
  geom_histogram(aes(y = ..density..), binwidth=0.2, fill="#0072B2", colour="black") +
  ggtitle("Neg Relationship Events > 0") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
# Create new variable for categorical transformation of trait self-esteem
dehart$rosn_cat <- cut(dehart$rosn, breaks=c(-1, 2.8, 3.4, Inf), labels = c("low", "mid", "high"))
```

Our main output variable of number of drinks consumed is highly positively skewed. Around 60% of the data is spread evenly across 0, 1, or 2 drinks, and only around 6% of the data is at or above 7 drinks. In

particular, the five data points with over 12 drinks may be high leverage, an issue we should look for in our modeling. Our other potential output variable of number of desire to drink does not display this extreme skew. Instead, it looks a bit more normal with a higher density area from 4 to 6, and tails of fairly uniform lower density out to 1 and 8.

Our main independent variable of interest in analyzing our current hypothesis is the number of negative relationship events. This is highly positively skewed, with 77% of our data points having a vlue of 0. Among the remaining non-zero data, we still see a positive skew with most of the data having values less than 1, and only a couple points with values over 3. There is a particular outlier at a value of 8 which may have high leverage that we should look for during our modeling. This lack of variation in our primary independent variable of interest may make our analysis more challenging.

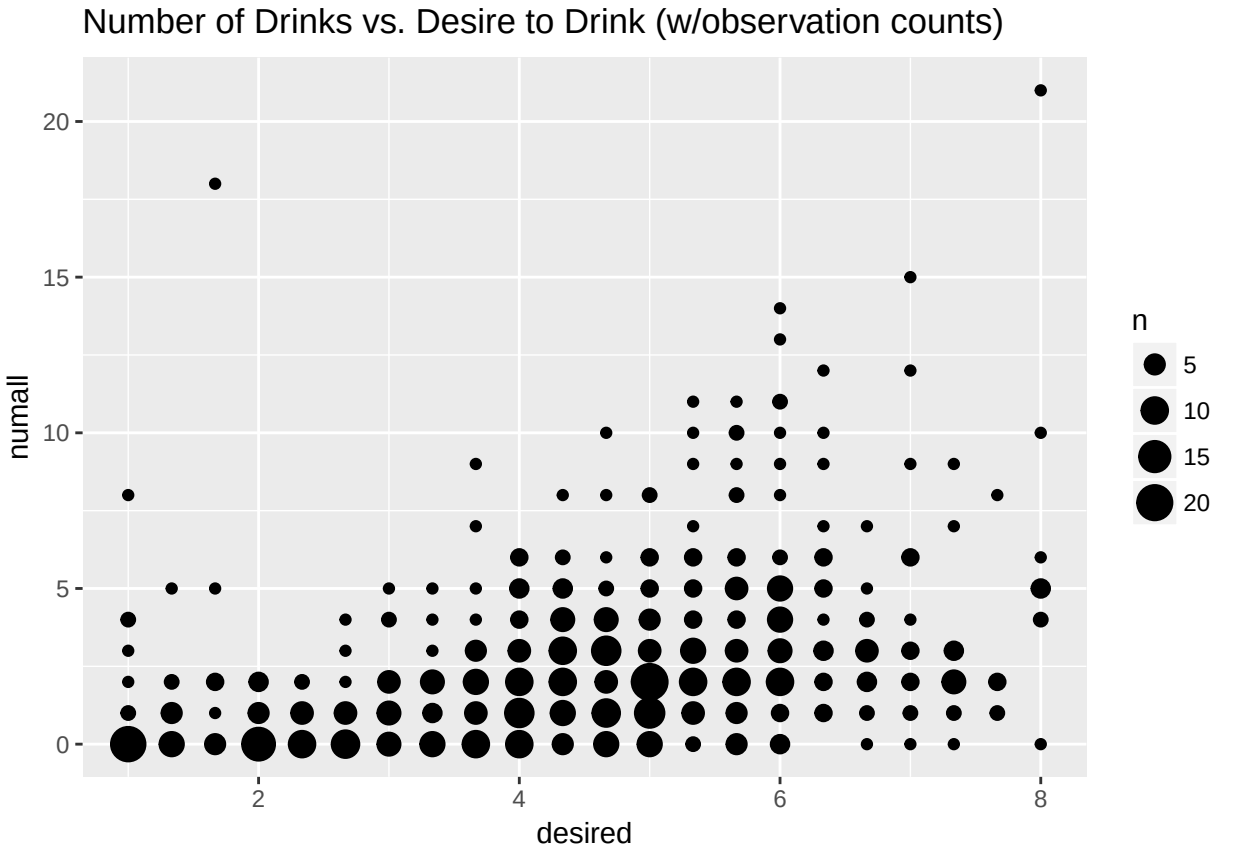
Our other variables measuring negative / positive event indexes also tend to be positively skewed, although less severely than negative relationship events. The positive measures have a larger spread than the negative ones - ie, positive romantic events is much less skewed than negative romantic events, and positive overall events has data going slightly above 3 while negative overall events maxes out less than 2.

Trait and state self-esteem are both negatively skewed. State self-esteem looks a bit closer to a normal distribution, likely due to the increased number of unique observations since this is a daily measure. Trait self-esteem is another primary independent variable since we will be looking for different behavior among those with high vs low trait self-esteem. There is not a clear break-point in the distribution that would easily divide the population into low vs. high self-esteem populations. This would indicate that we may want to include this as a continuous variable in our modeling rather than transforming it into a categorical variable. If we do consider transforming it, we do see slight jumps in density at 2.8 and 3.4, so would suggest 'low self-esteem' to be below 2.8, 'medium self-esteem' to be 2.8 through 3.4, and 'high self-esteem' to be above 3.4.

Age is relatively uniformly spread from mid-twenties through early forties, with slightly more participants on the older end of the spectrum.

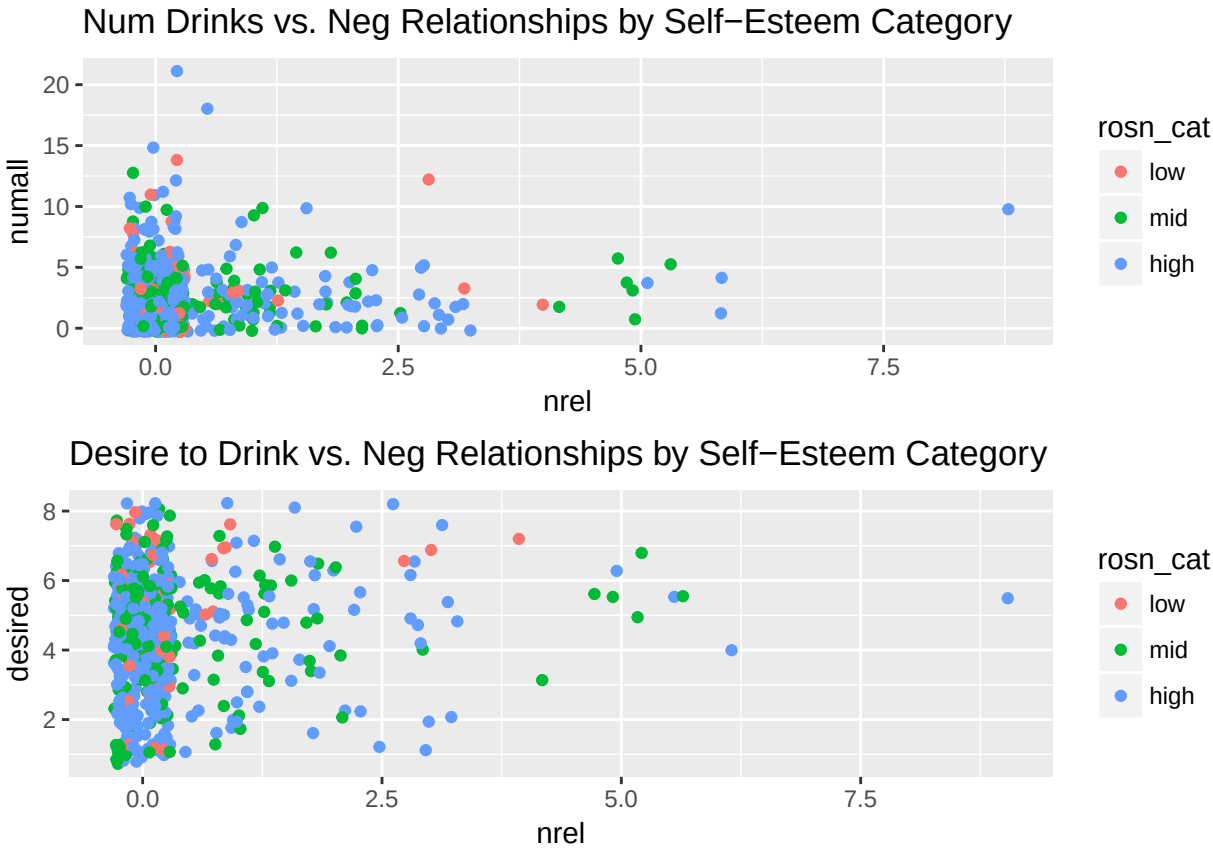
Having examined our univariate distributions, we will now consider multi-variate relationships.

```
# Number of drinks and desire to drink
ggplot(na.omit(dehart), aes(x=desired, y=numall)) +
  geom_point() + geom_count() +
  ggtitle("Number of Drinks vs. Desire to Drink (w/observation counts)")
```

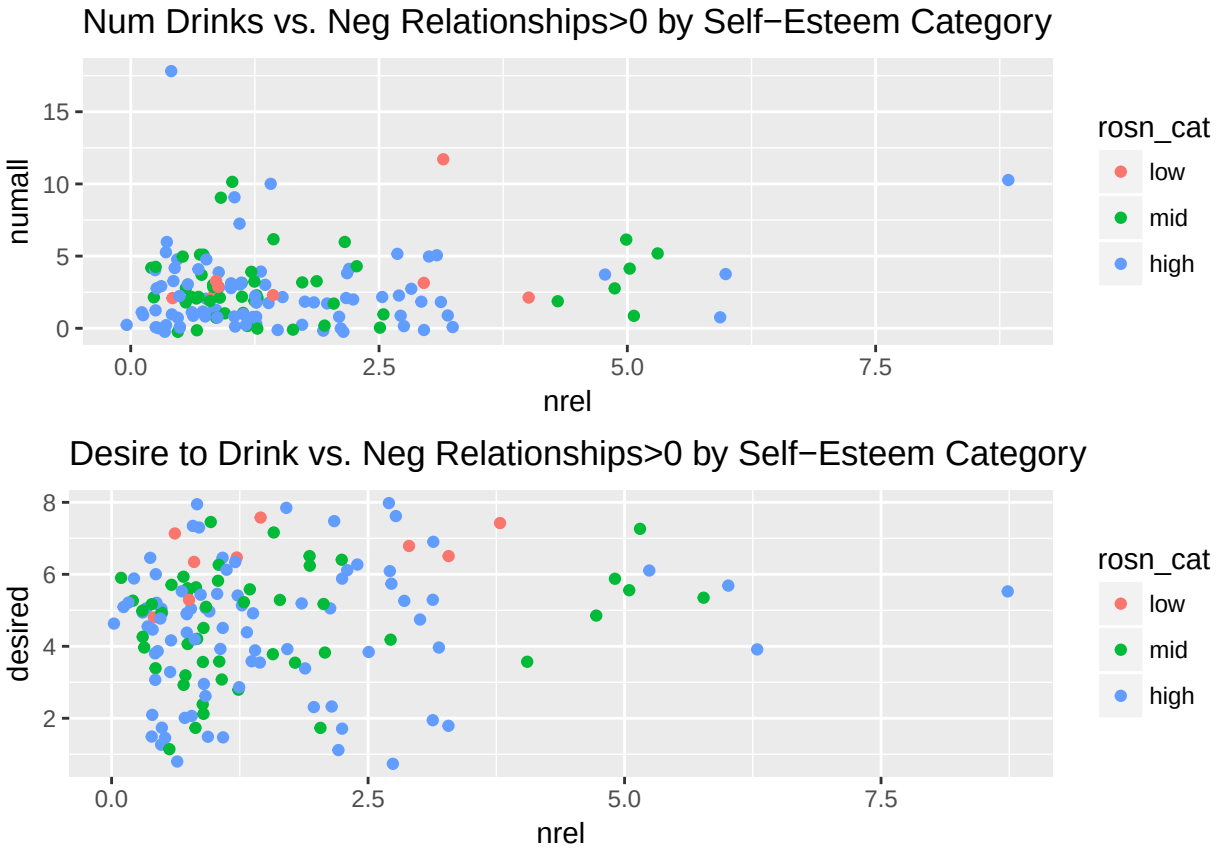


We see a generally positive relationship between desire to drink and number of drinks consumed, as we would expect. There are a couple outliers at the low end of desire to drink, but with a higher number of drinks consumed (even a point with 18 drinks).

```
# Drinking outputs vs. negative romantic relationship events, marking trait self-esteem category
g1 <- ggplot(na.omit(dehart), aes(x=nrel, y=numall, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Num Drinks vs. Neg Relationships by Self-Esteem Category")
g2 <- ggplot(na.omit(dehart), aes(x=nrel, y=desired, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Desire to Drink vs. Neg Relationships by Self-Esteem Category")

grid.arrange(g1, g2, nrow=2, ncol=1)
```



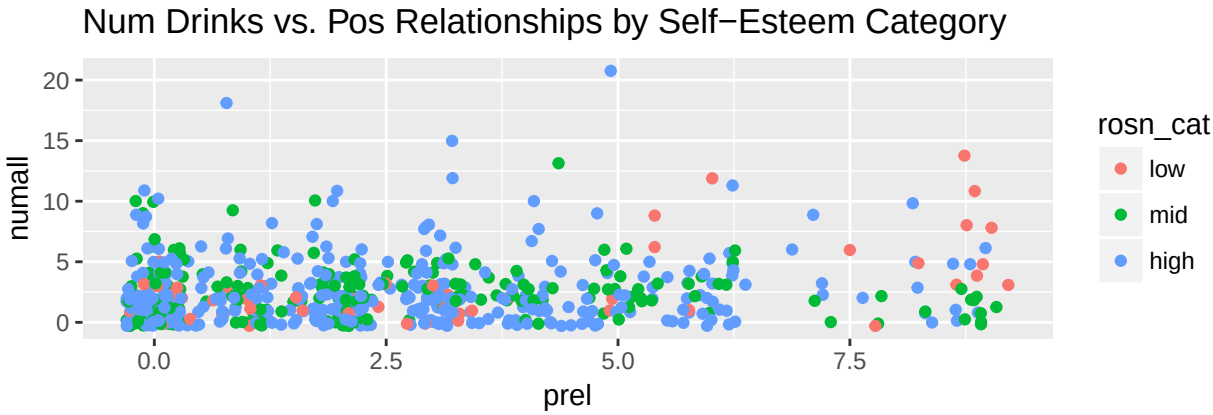

```
g1 <- ggplot(na.omit(dehart[dehart$nrel>0,]), aes(x=nrel, y=numall, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Num Drinks vs. Neg Relationships>0 by Self-Esteem Category")
g2 <- ggplot(na.omit(dehart[dehart$nrel>0,]), aes(x=nrel, y=desired, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Desire to Drink vs. Neg Relationships>0 by Self-Esteem Category")

grid.arrange(g1, g2, nrow=2, ncol=1)
```



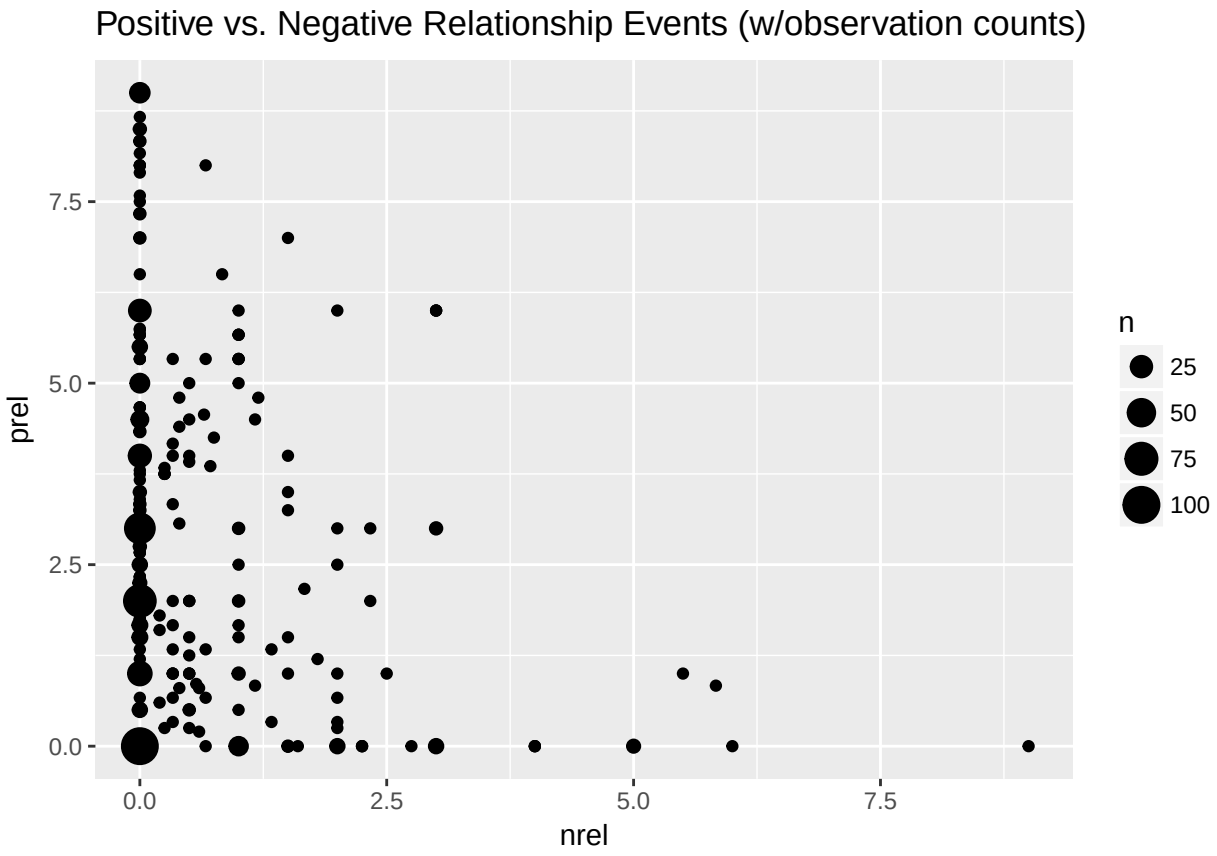
Here we look at all our primary variables - outputs, main independent variable of interest, and the groups where we expect to see different impacts of the independent variable. We see that negative relationship events do not appear to have a strong relationship with either number of drinks consumed or desire to drink. The large proportion of individuals with 0 negative relationship events easily covers the full range of the output variables, and we do not see major differences in spread for those with more negative relationship events. The one thing of note is that after around 3 for negative relationship events, we no longer see individuals consuming 0 drinks and their desire to drink is always above 3. However, this should be taken with a grain of salt since there are not many observations with more than a 3 for negative relationship events.

However, this relationship appears to change when we take into account long-term self-esteem (shown as categories here for ease of visualization). Those with low self-esteem and no negative romantic relationship events can be seen along the full spread of number of drinks consumed and desire to drink. But all low self-esteem individuals with some negative romantic relationship events consumed over around 2.5 drinks and rated their desire to drink over around 5. This is in contrast to those with high self-esteem who continue having lower number of drinks and desire to drink even in the presence of negative relationship events. These differences are more clearly visible in the scatterplots without the data points for those with 0 negative relationship events.

```
# Drinking outputs vs. positive romantic relationship events, marking trait self-esteem category
g1 <- ggplot(na.omit(dehart), aes(x=prel, y=numall, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Num Drinks vs. Pos Relationships by Self-Esteem Category")
g2 <- ggplot(na.omit(dehart), aes(x=prel, y=desired, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Desire to Drink vs. Pos Relationships by Self-Esteem Category")

grid.arrange(g1, g2, nrow=2, ncol=1)
```



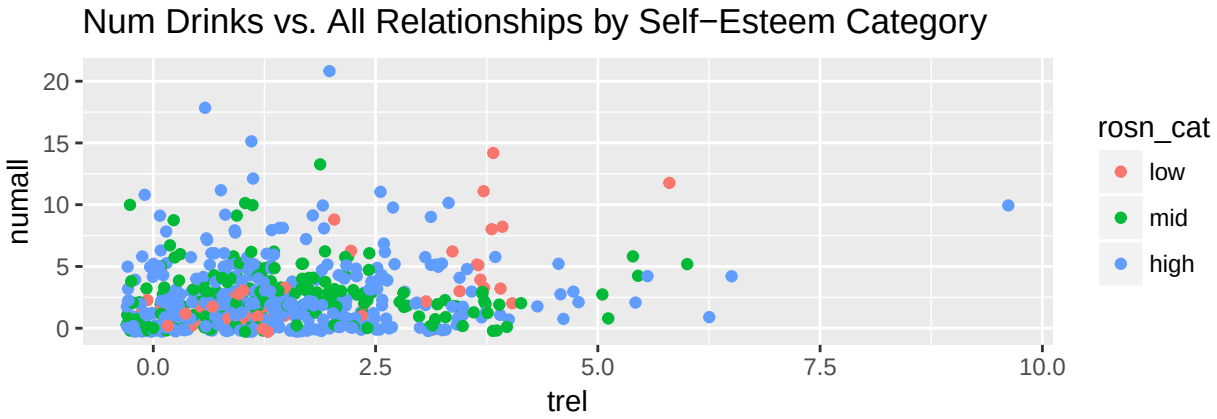

```
# Relationship between negative and positive relationship events
ggplot(na.omit(dehart), aes(x=nrel, y=prel)) +
  geom_point() + geom_count() +
  ggtitle("Positive vs. Negative Relationship Events (w/observation counts)")
```

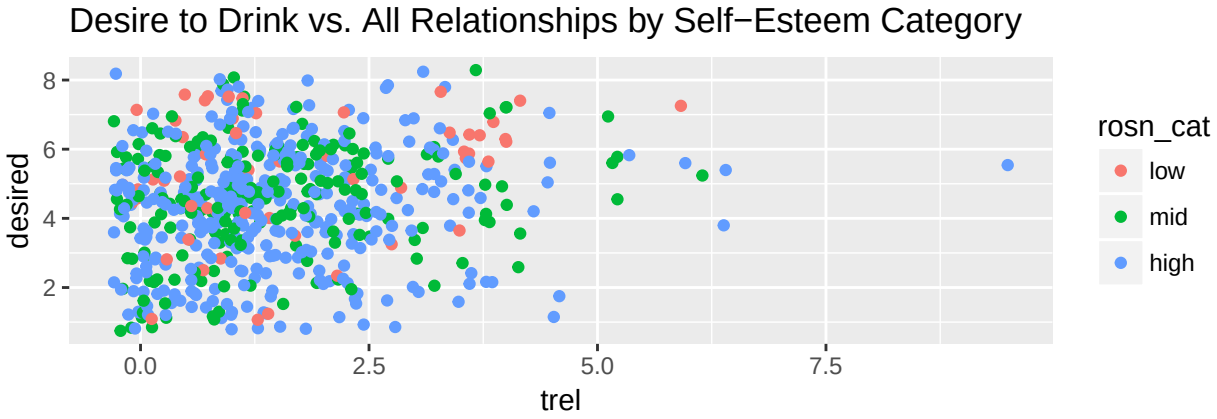


Running the same visualization for positive relationship events instead of negative ones, we similarly see the drinking outcome variables having similar distributions conditional only on our index of positive relationship events. As with negative events, we see a small relationship where those with more positive relationship events no longer report the lowest values for desire to drink (but less of a relationship for drinks consumed). When we take into account trait self-esteem levels, we also see different patterns emerge. Among low self-esteem individuals, those with higher values for positive relationship events also have the higher number of drinks consumed and larger concentration at a high desire to drink. Those with high self-esteem do not display a similar trend, and appear to have less change in their distributions based on number of positive romantic events. Combined with the patterns seen for negative relationship events, this may mean that low income individuals have a higher desire and consumption of alcohol when having more romantic relationship events, regardless of whether they are positive or negative.

It appears that there is a modestly negative relationship between positive and negative romantic relationship events, but it is clouded by the large proportion of observations with 0 negative events.

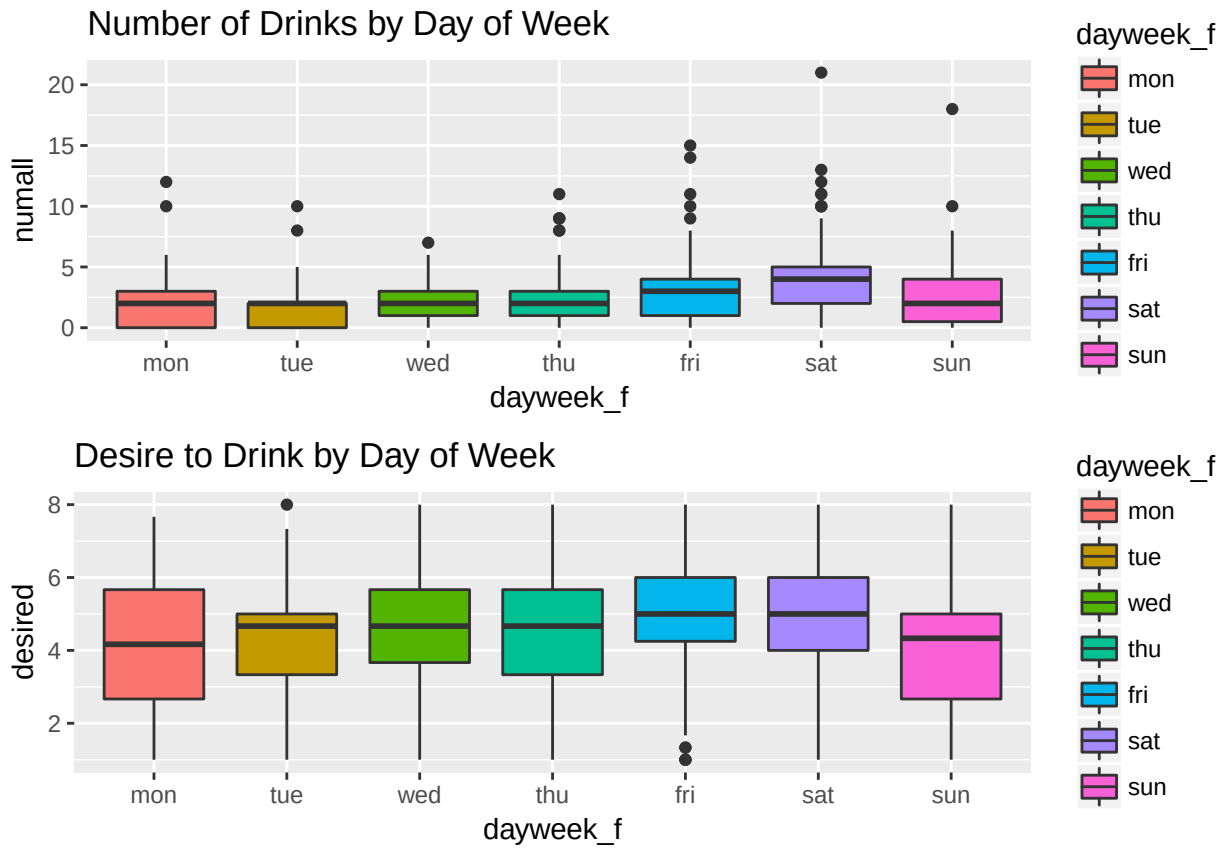
```
dehart$trel <- dehart$nrel/sd(dehart$nrel) + dehart$prel/sd(dehart$prel)

# Drinking outputs vs. total romantic relationship events, marking trait self-esteem category
g1 <- ggplot(na.omit(dehart), aes(x=trel, y=numall, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Num Drinks vs. All Relationships by Self-Esteem Category")
g2 <- ggplot(na.omit(dehart), aes(x=trel, y=desired, color=rosn_cat)) +
  geom_jitter(width=0.3, height=0.3) +
  ggtitle("Desire to Drink vs. All Relationships by Self-Esteem Category")

grid.arrange(g1, g2, nrow=2, ncol=1)
```



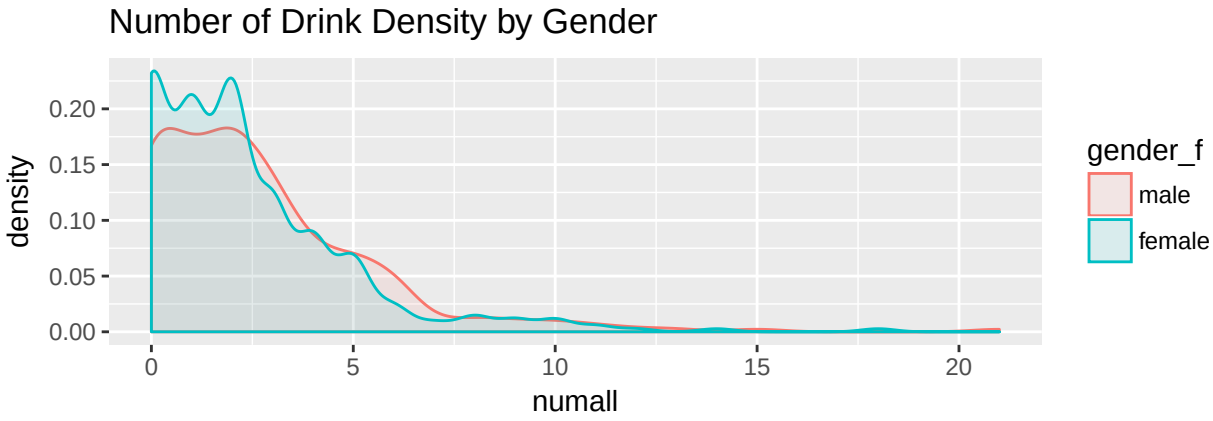


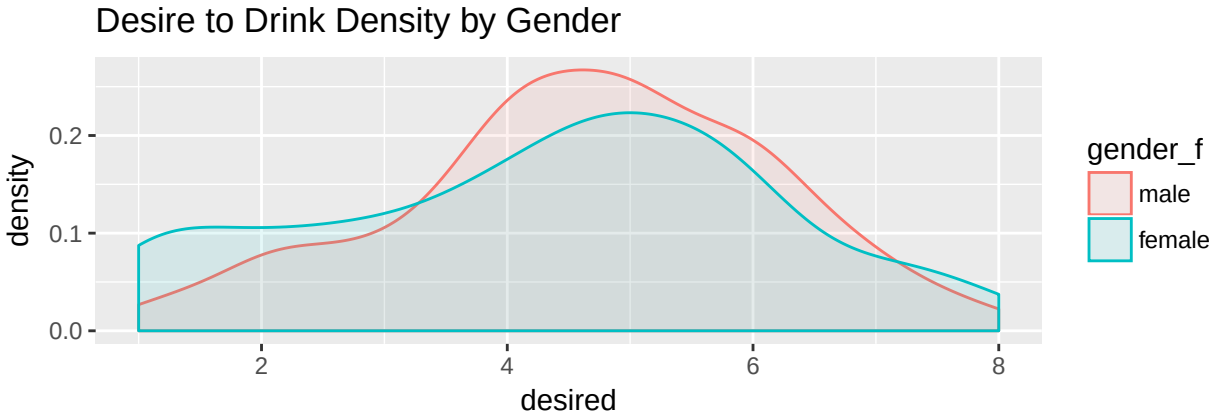
I combined the two by adding the number of standard deviations to avoid the negative events being overwhelmed by the positive ones since so many of the negative values were close to 0 and positive ones had a wider spread. If these had been actual counts of interactions, I would have just added them. We see the same sort of relationships here as we saw in both the negative and positive relationship cases.

```
# Boxplots of drinking outcomes by day of the week
g1 <- ggplot(na.omit(dehart), aes(dayweek_f, numall)) +
  geom_boxplot(aes(fill = dayweek_f)) +
  ggtitle("Number of Drinks by Day of Week")
g2 <- ggplot(na.omit(dehart), aes(dayweek_f, desired)) +
  geom_boxplot(aes(fill = dayweek_f)) +
  ggtitle("Desire to Drink by Day of Week")

grid.arrange(g1, g2, nrow=2, ncol=1)
```

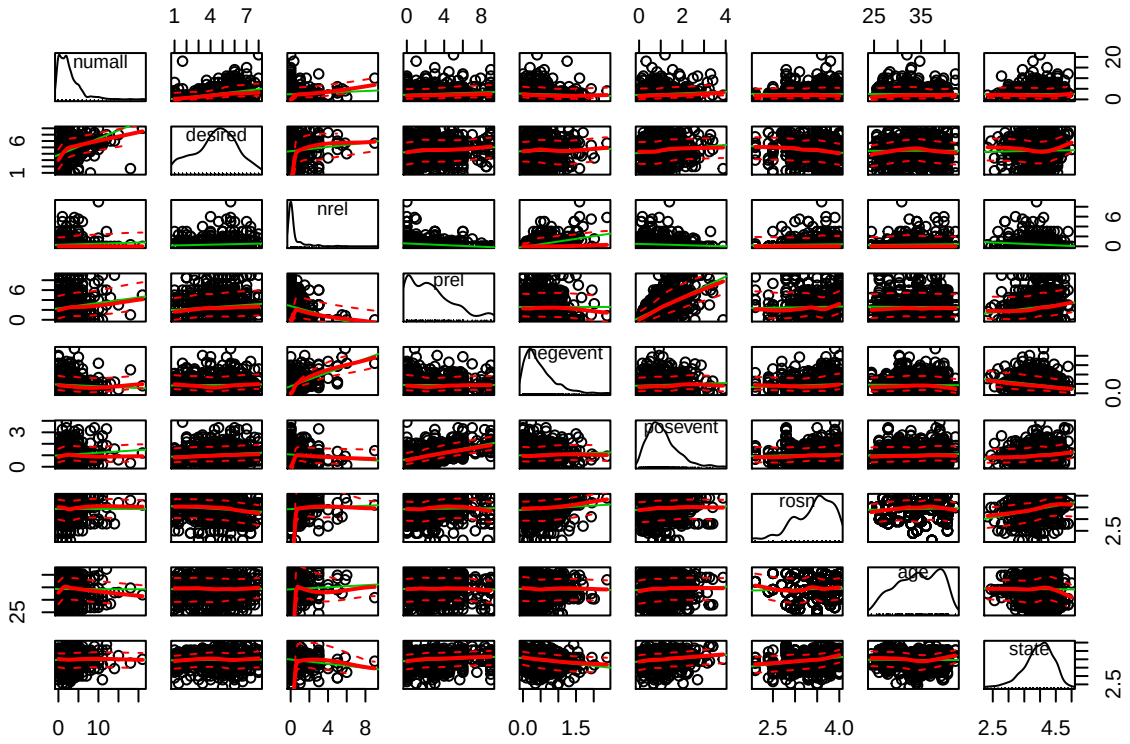



```
# Density of drinking outcomes by gender
g1 <- ggplot(na.omit(dehart), aes(numall)) +
  geom_density(aes(color=gender_f, fill=gender_f), alpha=0.1) +
  ggtitle("Number of Drink Density by Gender")
g2 <- ggplot(na.omit(dehart), aes(desired)) +
  geom_density(aes(color=gender_f, fill=gender_f), alpha=0.1) +
  ggtitle("Desire to Drink Density by Gender")

grid.arrange(g1, g2, nrow=2, ncol=1)
```






```
# Scatterplot matrix of all numerical variables
suppressWarnings(scatterplotMatrix(na.omit(dehart[,c("numall", "desired", "nrel", "prel", "negevent", "
```



Beyond our main relationships, we also wanted to look at the other covariates' relationships with our output variables. For day of the week, we see moderately increased number of drinks consumed and desire to drink on Friday and Saturday. The other days of the week have slightly lower means, with more variance on Monday and Sunday. For gender, we see that males have a higher density at mid to high values for desire to drink and their density for drinks consumed has a less steep slope going towards more drinks. Females, on the other hand, have higher densities at low desire to drink and low number of drinks consumed with steeper drops as drinking values increase. Note that we do not see large differences in relationships with covariates for desire to drink versus drinking consumption.

We also examined the scatterplot matrix of all the numerical variables for relationships on note. Beyond the ones we have already examined in more depth, we see that there are positive correlations between negative realtionship events and all negative events, as well as for positive relationship events and all positive events. We also see a moderately positive correlation between trait self-esteem and state self-esteem.

Questions for modeling raised by EDA

- How do we deal with the time series aspect of the data? Do we limit ourselves to 1 day, do we run 7 different models (one for each day of the week)? DO we include all data as if they are independent data points?
- Do we exclude the couple missing data points? If so, do we exclude all data for that participant as a whole?
- What is our output? Number of drinks, or desire to drink?
- If output is number of drinks, do we use a poisson model with number of drinks as the outcome, or do a ordinal model after binning the number of drinks into categories?
- Should we include trait self-esteem as a numerical variable, or decide on a split point for high vs low self-esteem?
- Do we want a variable for total romantic relationship events?
- Keep in mind checks for leverage on both number of drinks and number of negative relationship events.

Jeffrey's EDA