Homework 1

1. In order to clean the data, I did a few things. First, I changed the 0, 1 encoding of the sex variable to their appropriate genders. Then, I made this variable a factor. I scaled the income variable by 52, because it is measured per week. This allows the income and the gamble variable to be measured in the same scale. Finally, I divided the verbal variable by 12, since it is a score on a test out of 12 questions. This way the numbers represent something a little more intuitive. Here’s the code I used to complete this:

teengamb$sex[teengamb$sex == 0] <- "male"

teengamb$sex[teengamb$sex == 1] <- "female"

teengamb$sex <- as.factor(teengamb$sex)

teengamb$income <- teengamb$income \* 52

teengamb$verbal <- teengamb$verbal / 12

1. To complete this, I used the summary() function on each variable. For the sex variable, I received a count of each level, 19 female, 28 male.

summary(teengamb$sex)

The status variable has a minimum value of 18, a maximum value of 75, an IQR of 33.5. The median is 43, and the mean is 45.23. Since the mean is slightly greater than the median, this indicates our data is slightly right-skewed.

summary(teengamb$status)

The income variable has a minimum value of 31.2 and a maximum of 780 and the IQR 218.9. The median is 169 and the mean is 241.4. Like the status variable, this indicates that there is a right-skew in our data, though this one is more pronounced.

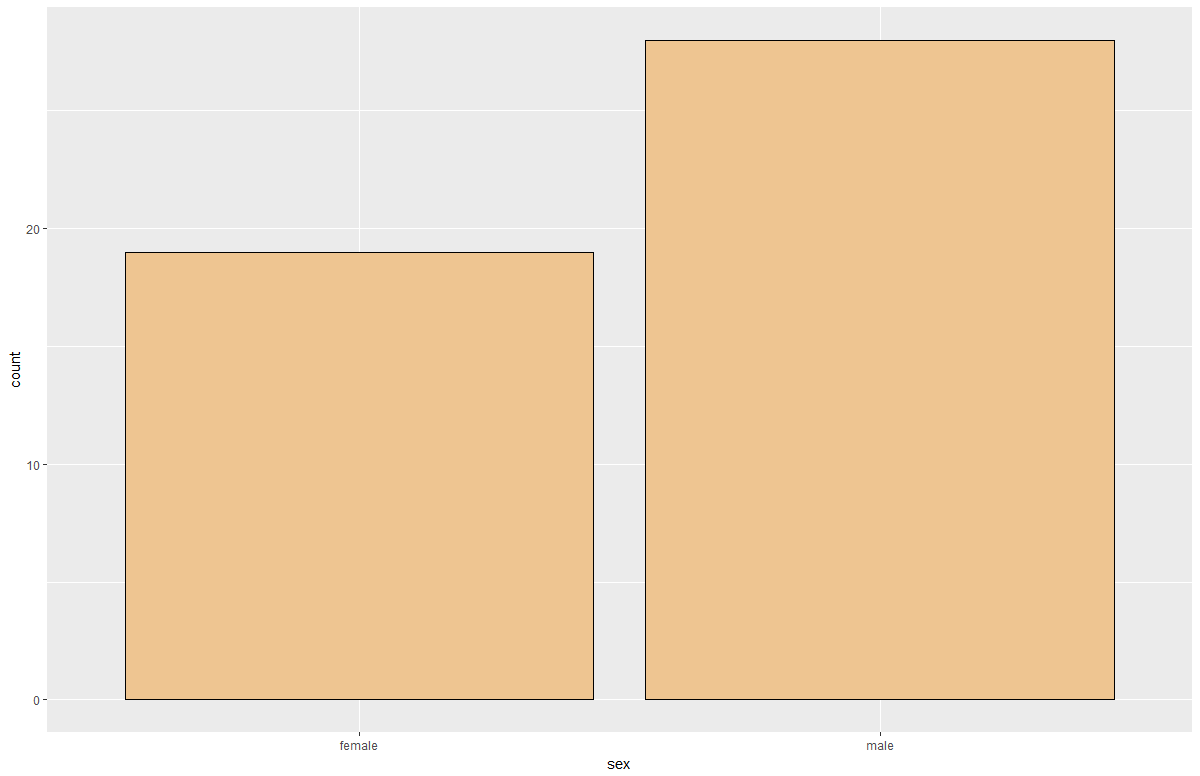
summary(teengamb$income)

The verbal variable has a minimum of .08333 and a maximum of .83330. The IQR is .16667, the median is .58330 and the mean is .55550. This indicates that there is a left-skew in our data. Not the place you want a left-skew!

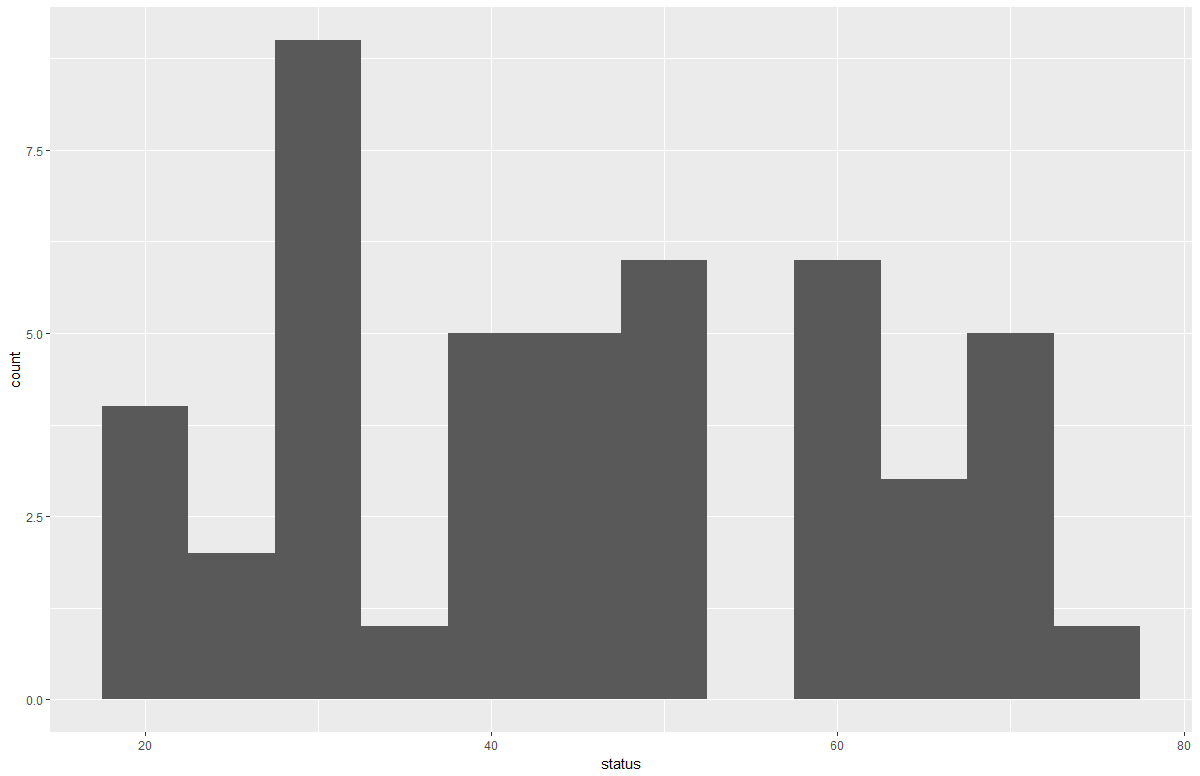
summary(teengamb$verbal)

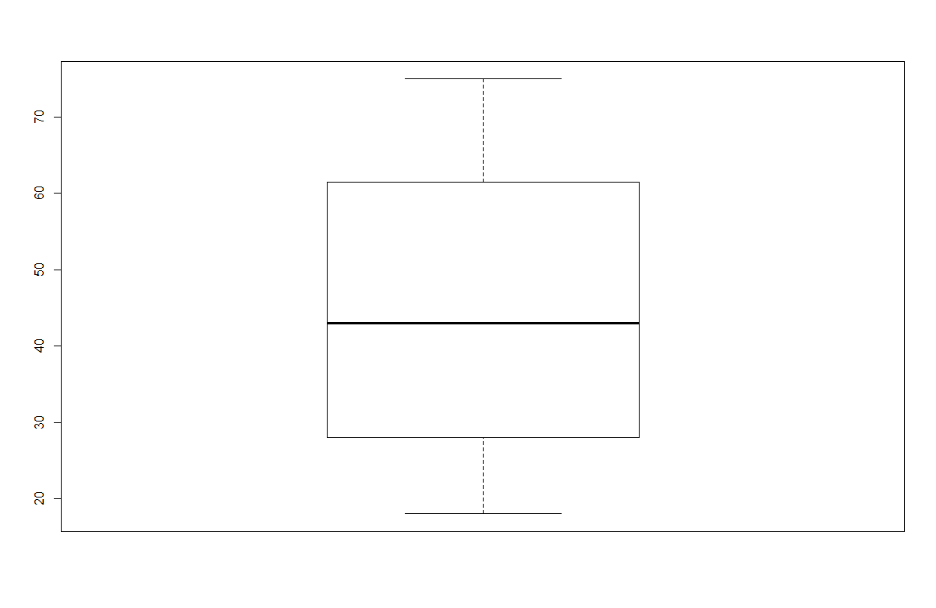
The gamble variable has a minimum of 0, a maximum of 156, and an IQR 18.3. This is by far the most interesting variable, because it is so heavily right-skewed. This can be seen by examining the median and mean, which have values of 6 and 19.3 respectively.

This simple bar chart depicts the counts of each gender in our data set. We can see from this that there are just under 20 females (19), and a little over 25 males (28).

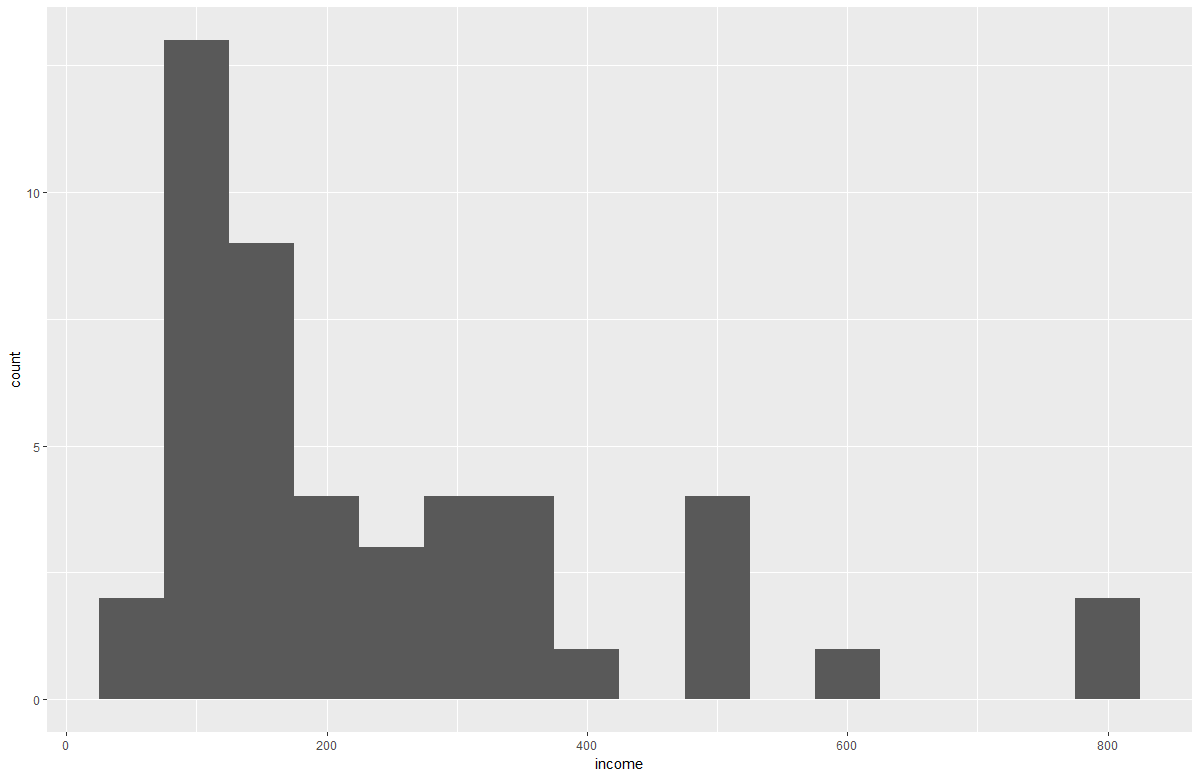


Here is a histogram and boxplot of the status variable. We can see from the histogram that the distribution is spread out, and the boxplot helps to visually confirm some of the information from the numerical summary in question 2.

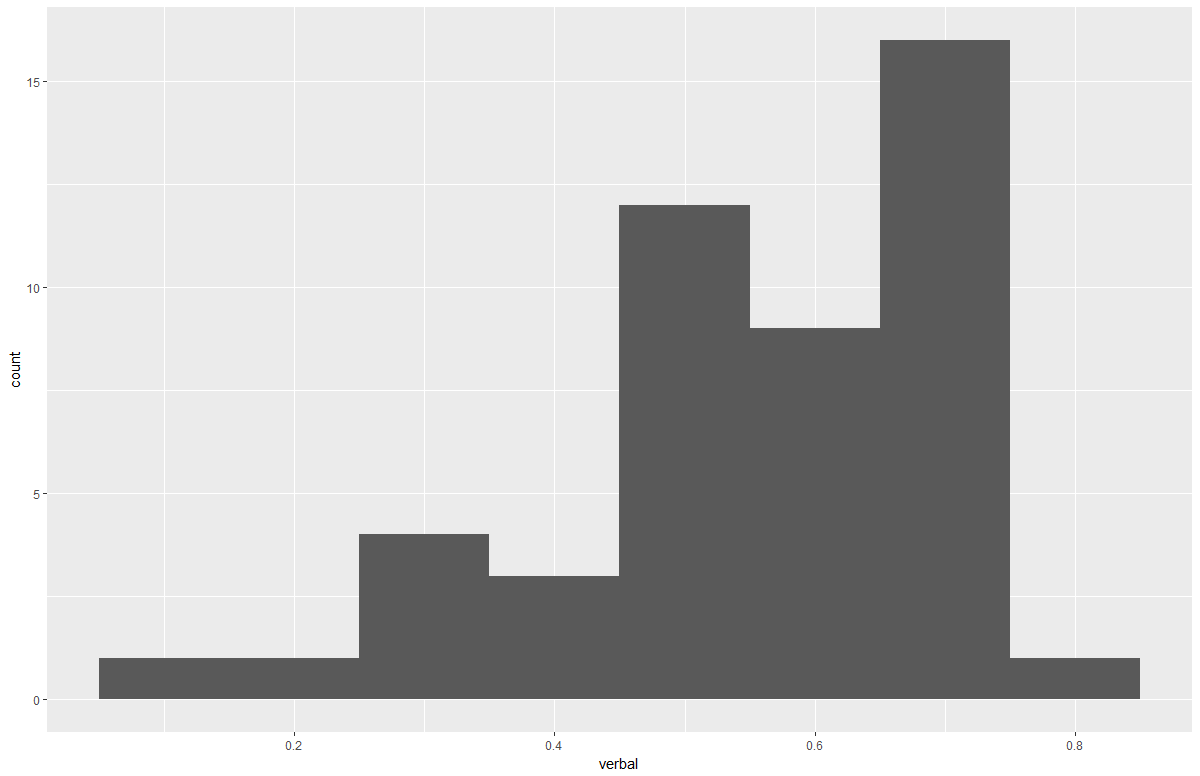




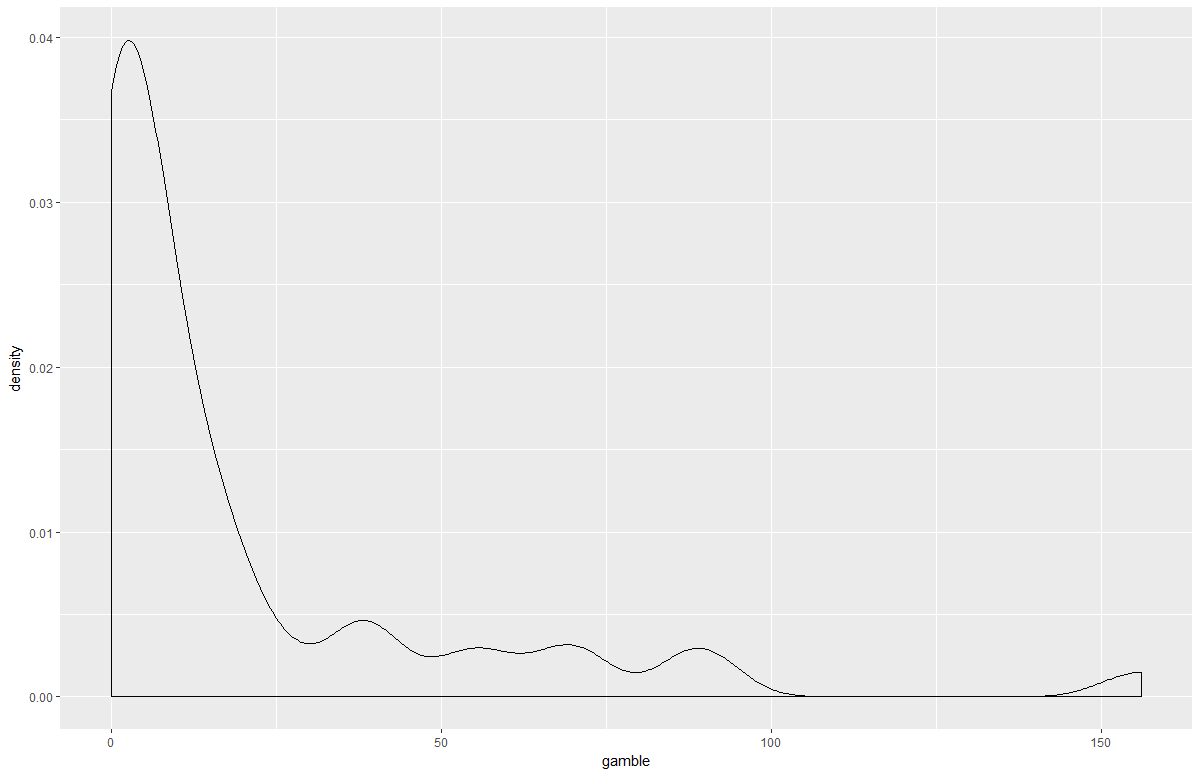
This is a histogram of the income variable. This helps to confirm that the data is right-skewed, which was mentioned in question 2.



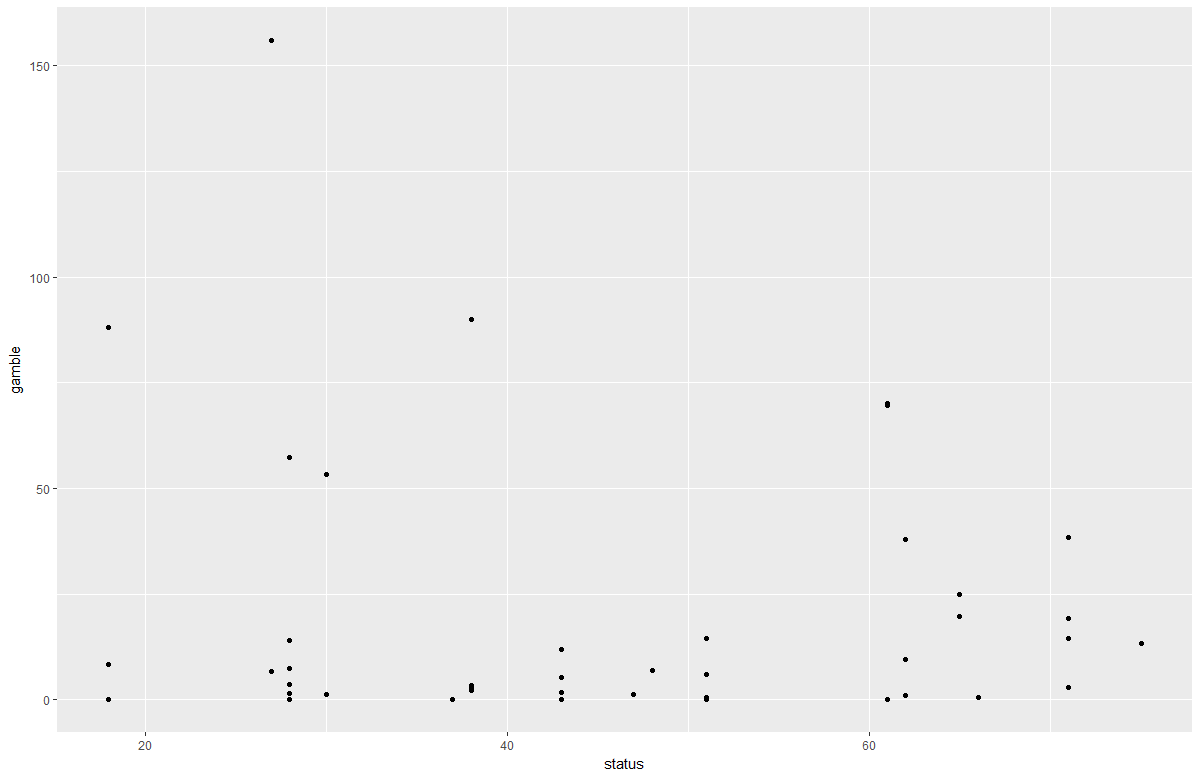
This is a histogram of the verbal variable. The majority of the scores are in the .5 - .8 range, but the data is still left-skewed.

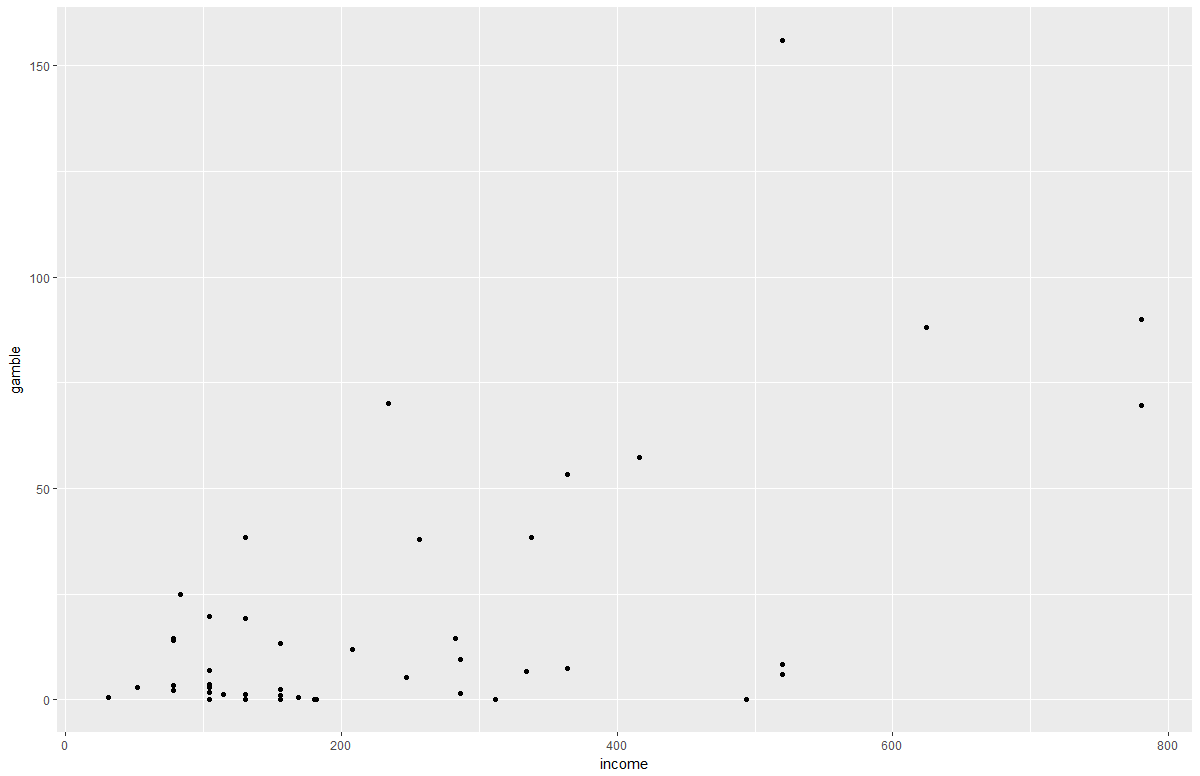


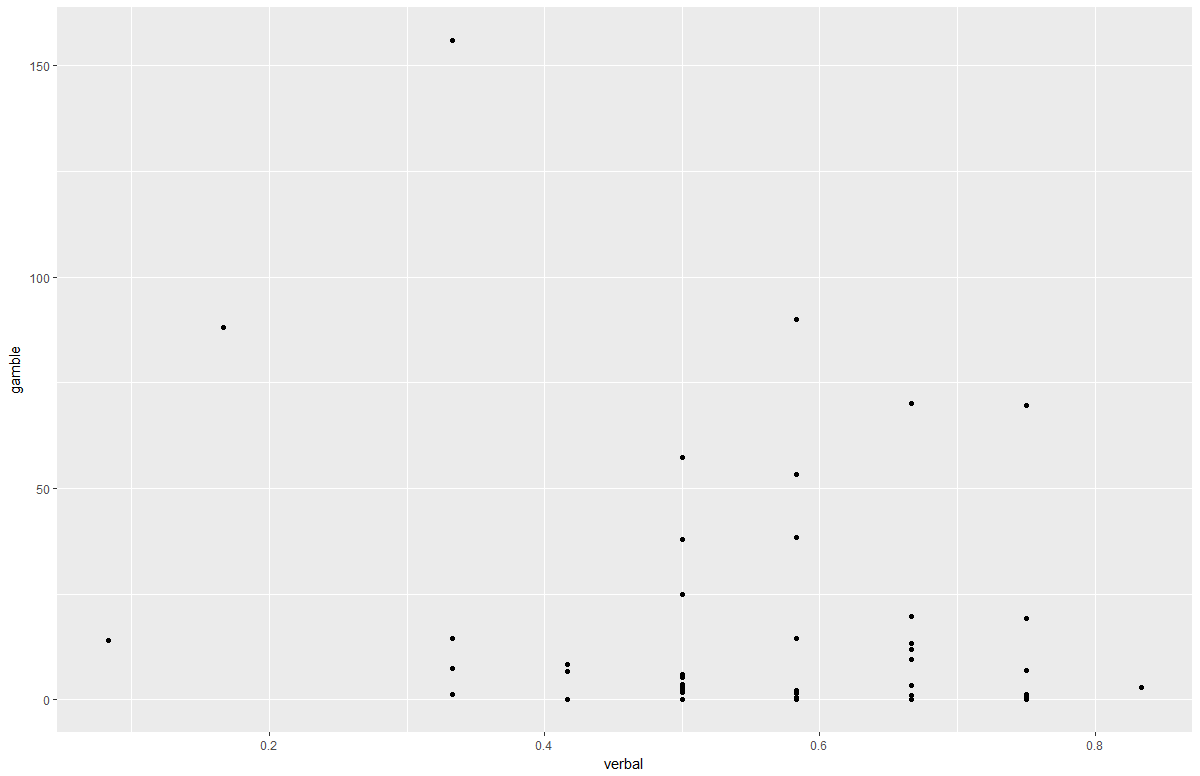
Just to change things up, here is a density plot for the gamble variable. Most of our gamble data is located between 0-25 pounds per year, but there is definitely a strong right-skew present, as evidenced by the tail.

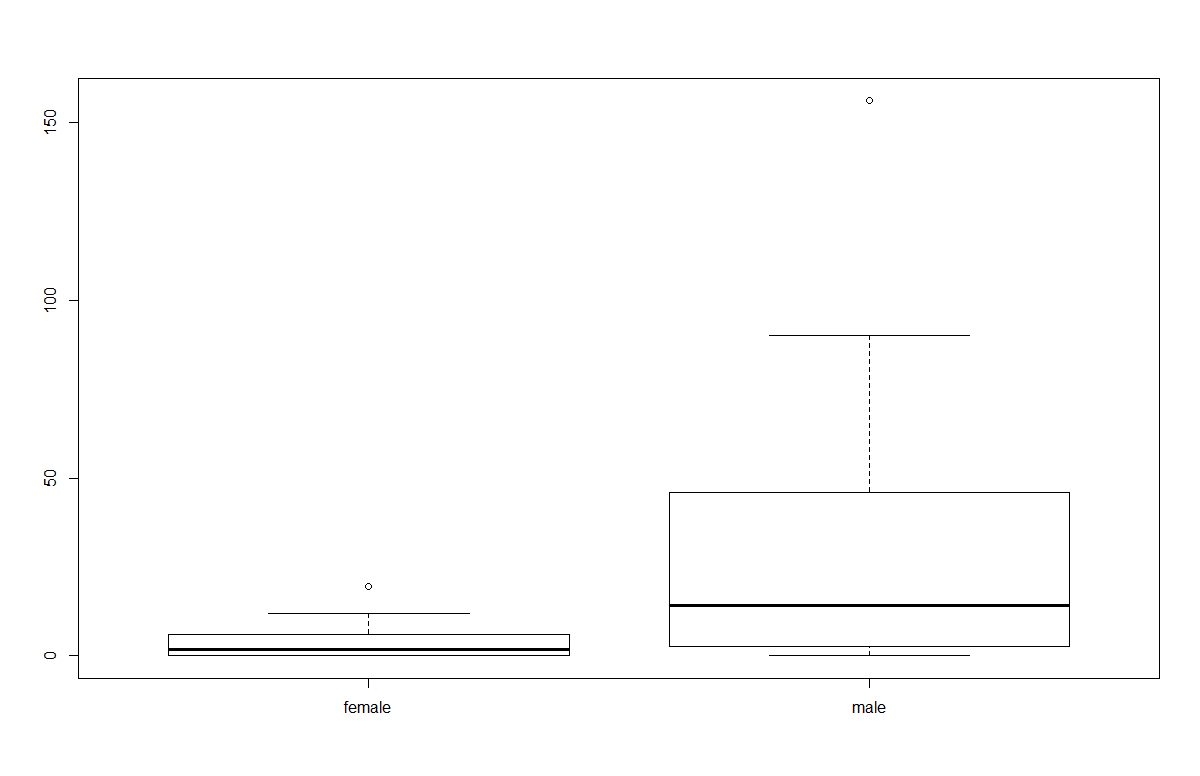


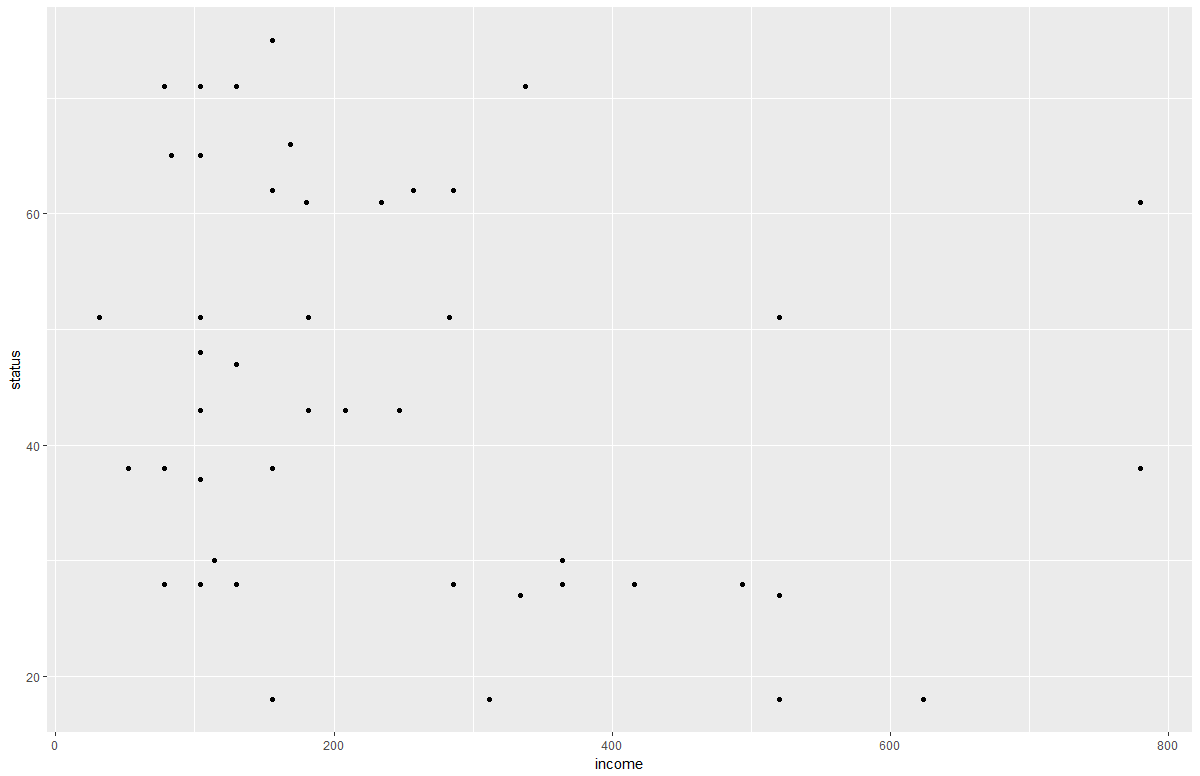
1. Here are the 10 bivariate plots of each variable compared with the others:

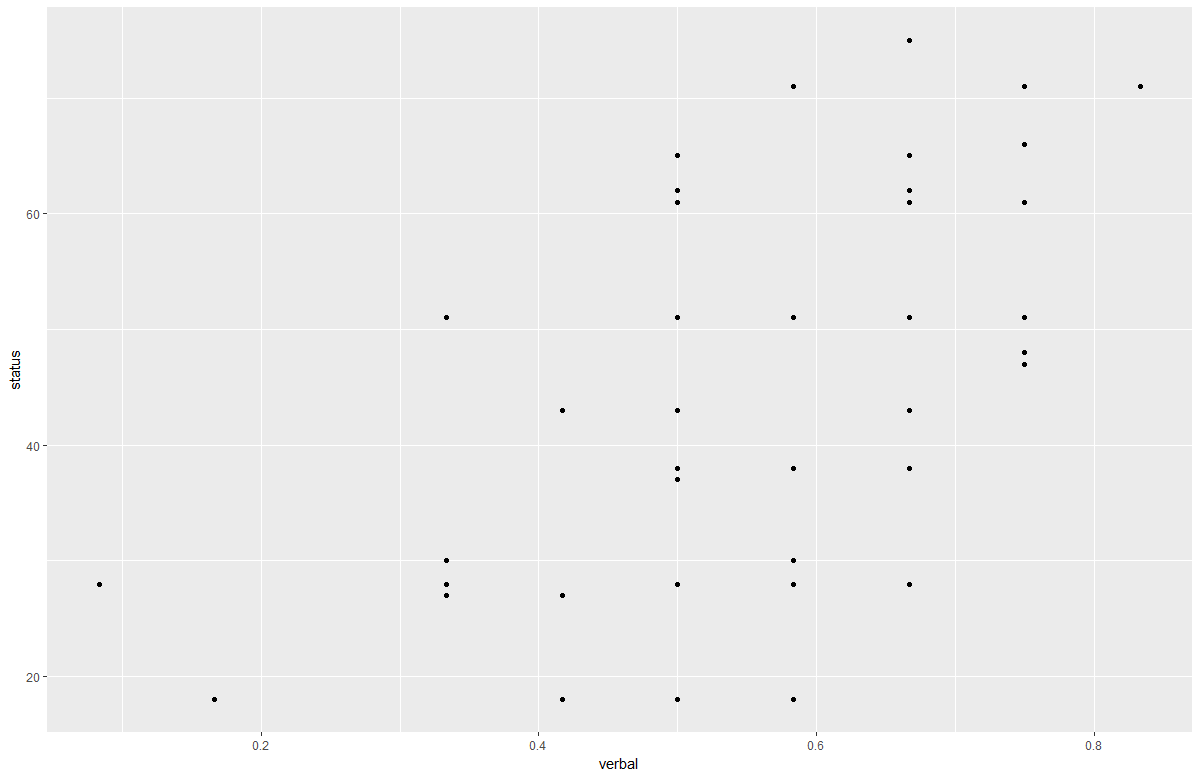


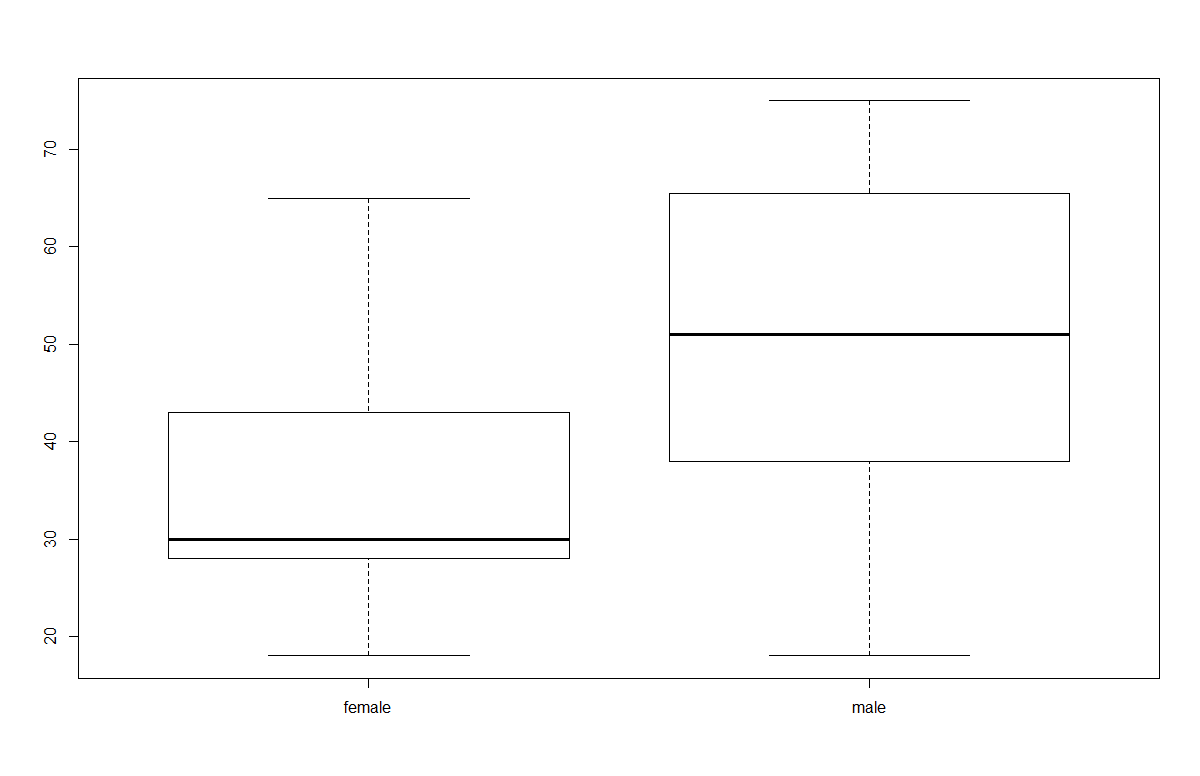


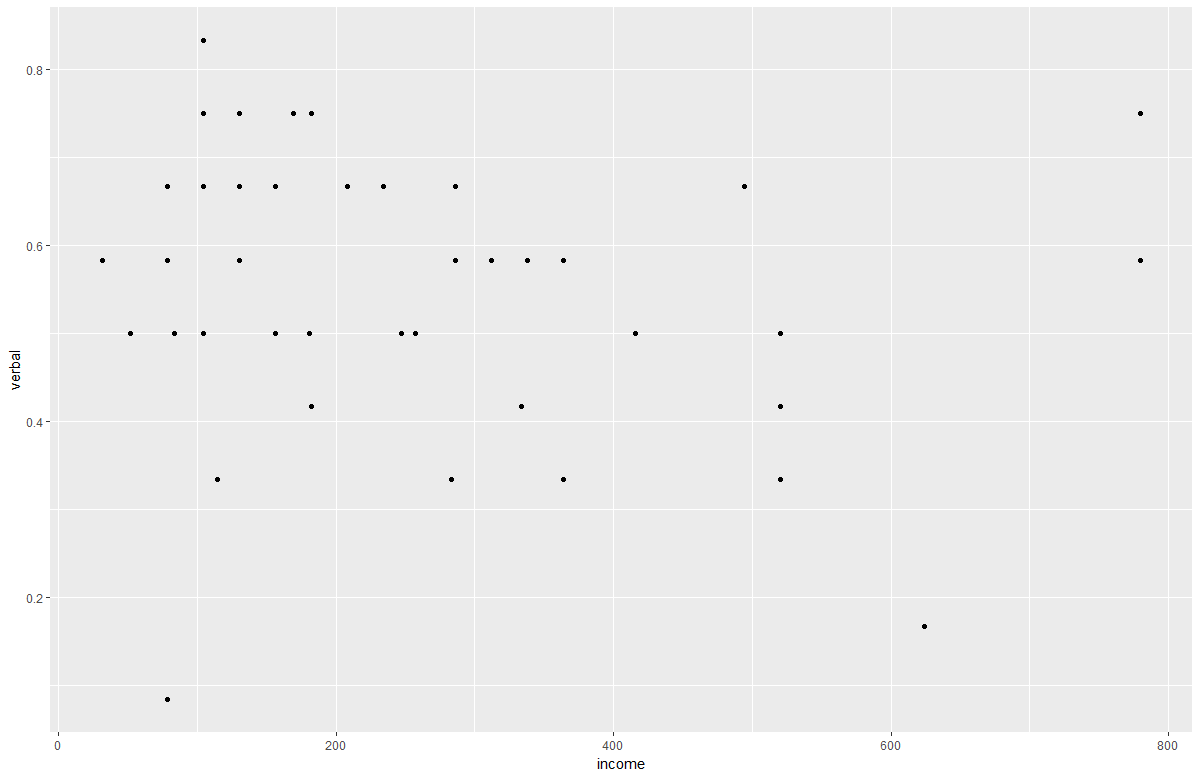


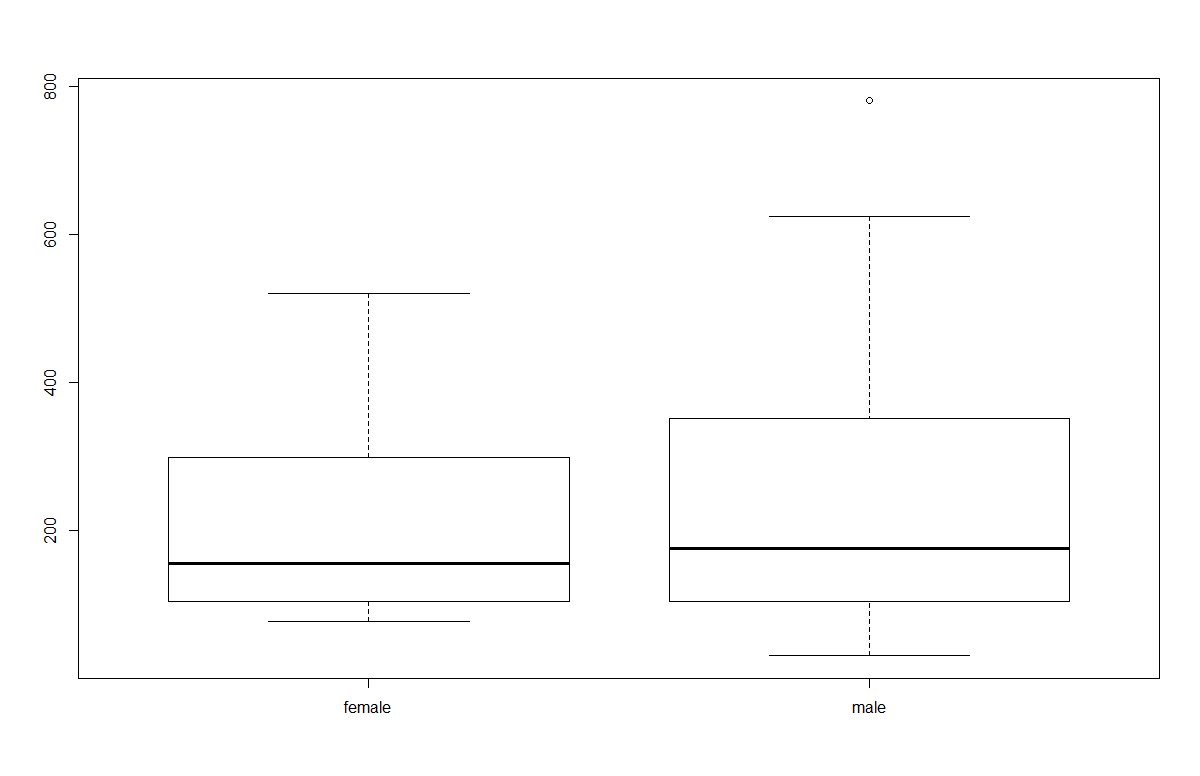
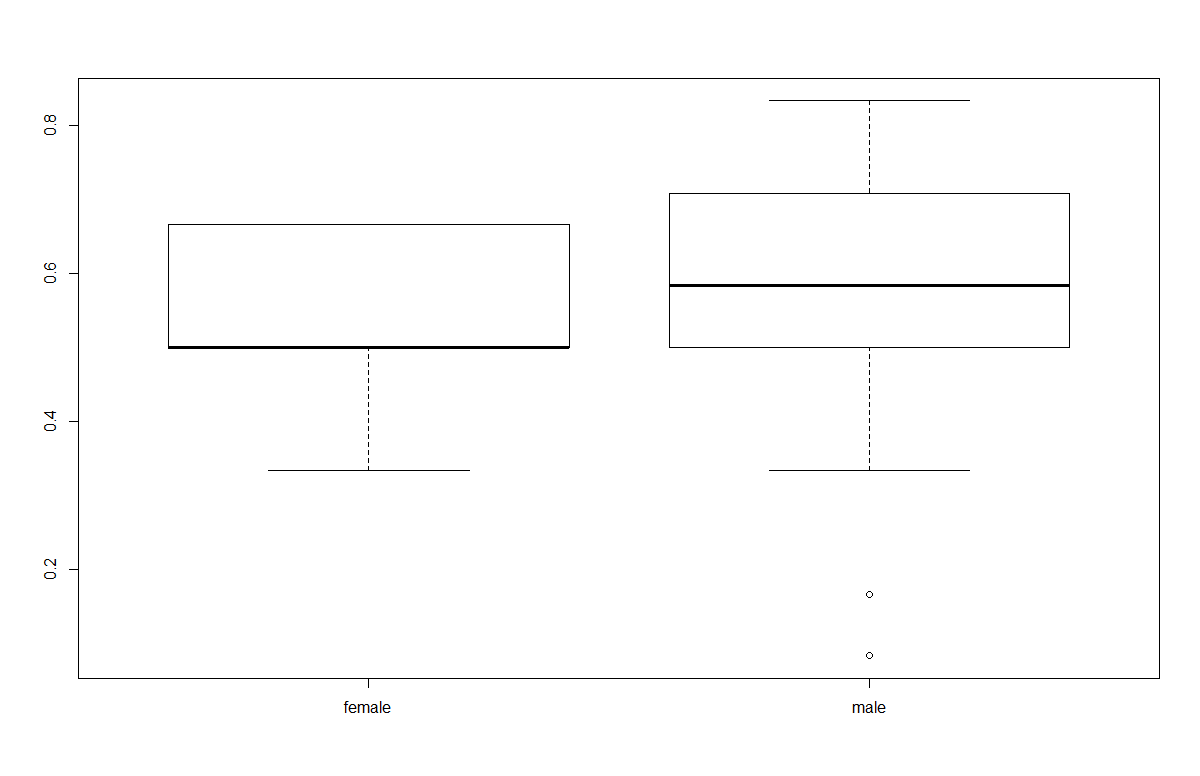






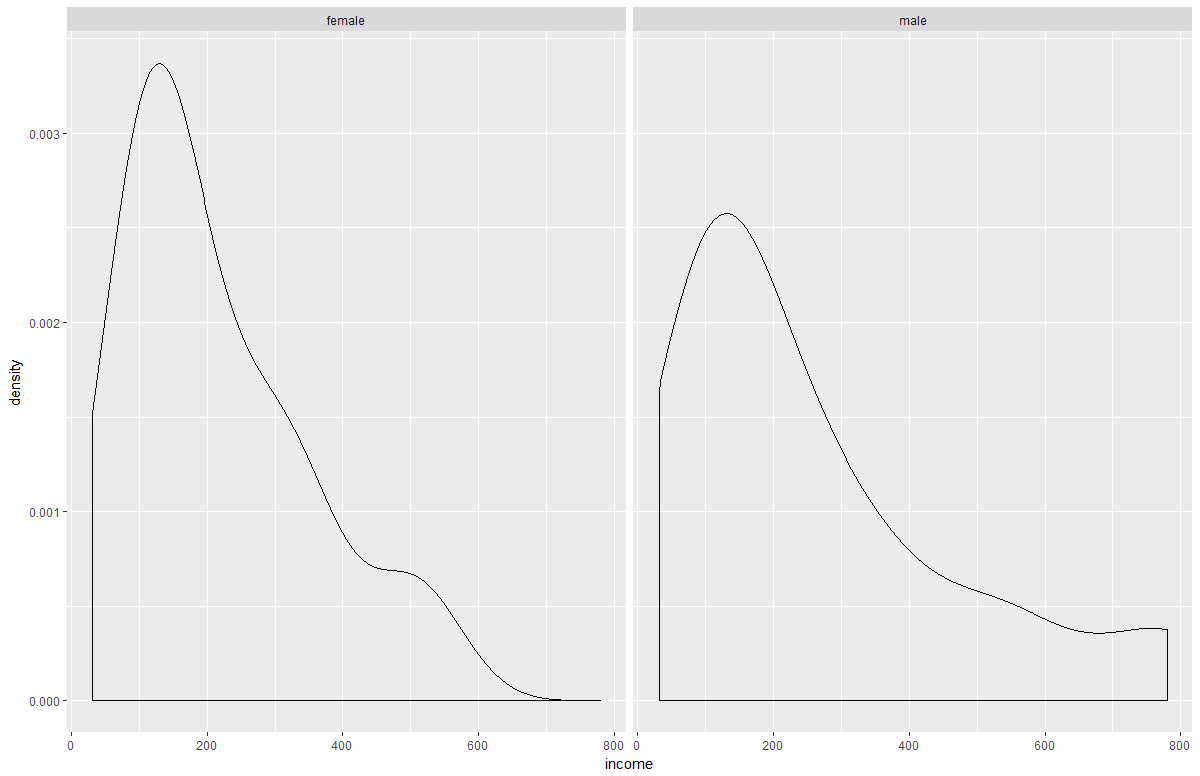




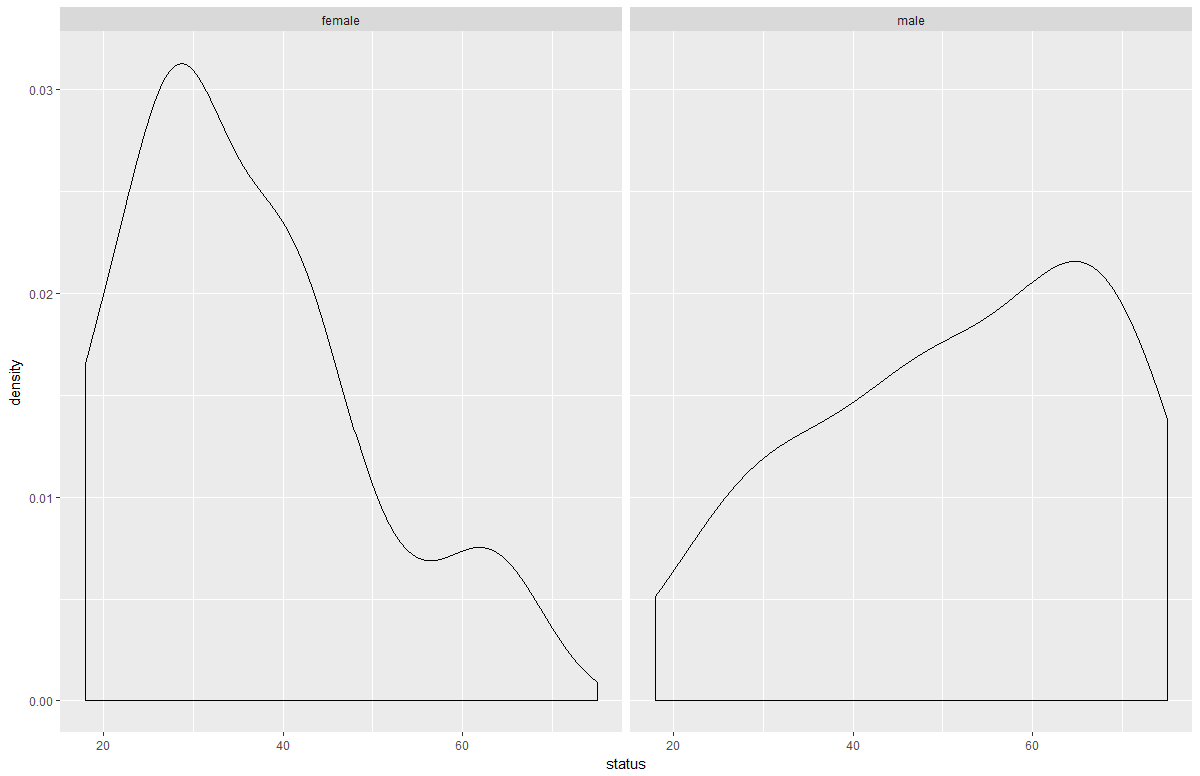


Some of the most interesting plots begin with the first split-level boxplot provided. This boxplot is for the gamble variable broken done by gender. One might believe that men gamble more than women, which is partially what this confirms. However, what I found most interesting is how much drastically spread out the boxplot is for the males versus the females. The females exhibit a much more controlled distribution, while makes exhibit one that is spread out and skewed. Another plot that is interesting to me is the status ~ income plot. I wouldn’t say that there is a very strong negative linear relationship, but there appears to be somewhat of a negative linear relationship. If I had to guess what the plot would look like without data, I would definitely have guessed that the more income someone makes, the higher the status of their parents were. The plot that appears to have the strongest relationship is the status ~ verbal plot. To me, this makes intuitive sense, because status observations are based on the subject’s parents. Thus, if your parents had higher status, then to me it makes intuitive sense that you would be more verbally capable. Another plot I found interesting was the verbal ~ income. I would have guessed that this would exhibit a strong positive relationship, but it certainly doesn’t.

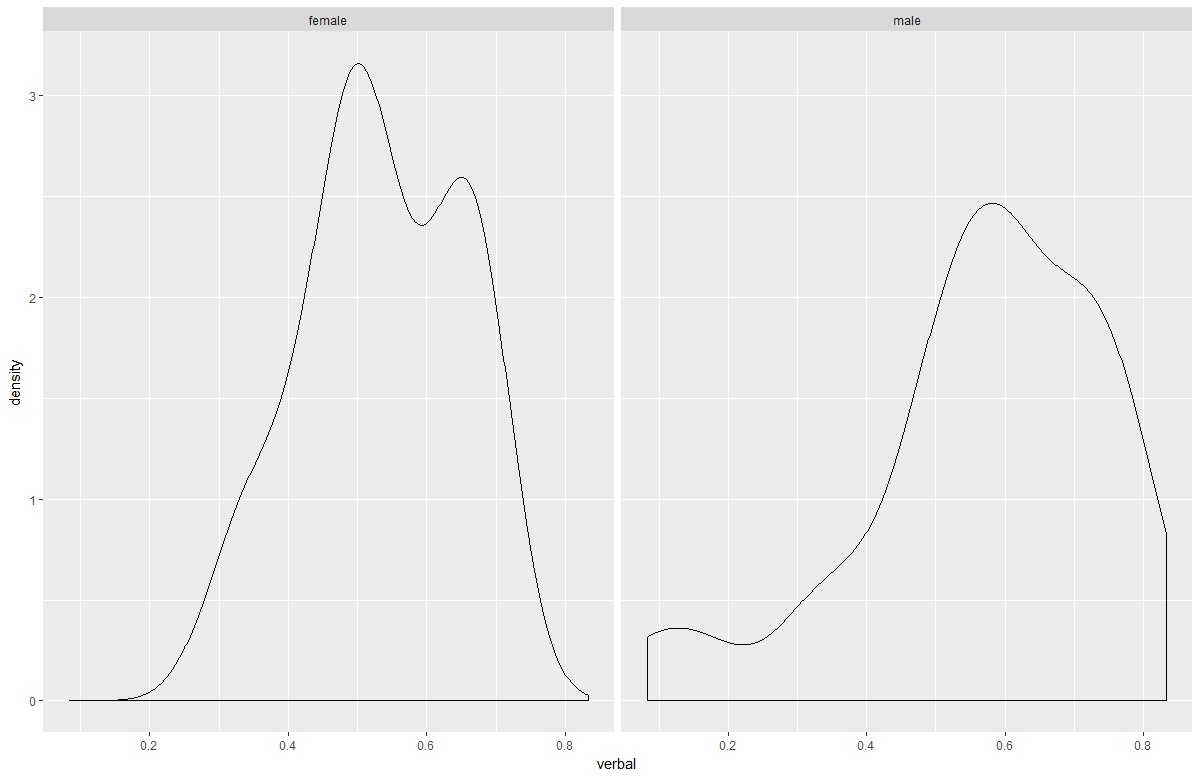
1. Here are density plots for the four quantitative variables faceted by sex:



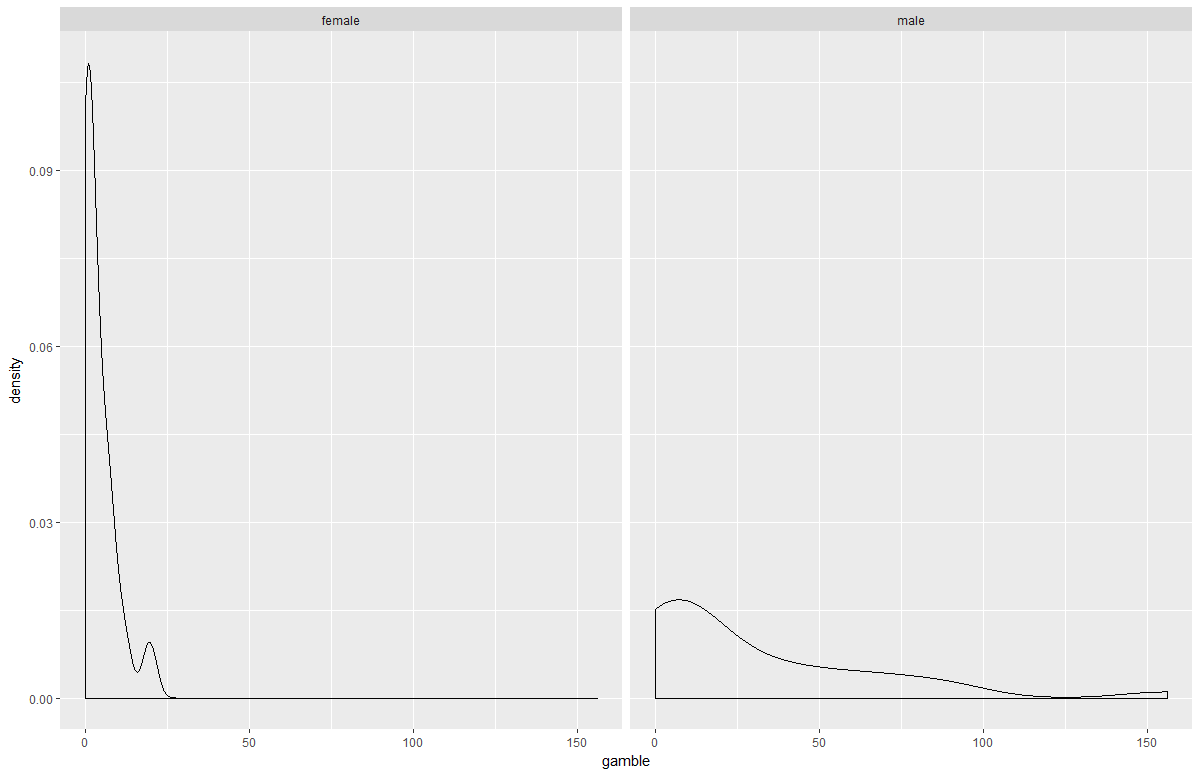
The income variable is more right-skewed for men than it is for women. Additionally, the women have a higher density around the 200 pound per year mark then the men.



The status variable is almost reversed based on sex. For the men, there is a higher density towards higher values of status, and vice versa for women. It’s hard to tell why this might be, especially because the status variable is representative of an observation’s parents.



The verbal scores are fairly similar, with more left-skew in the male scores, and slightly more density around 0.4-0.6 score range for women.



This is by far the most interesting density plot comparison. The women have a greater amount of density near low values of gambling (measured in dollars spent each year), while men have more right-skewed data. The differences in these densities is alarming, and would certainly help to draw some conclusions about people and their gambling habits.