# Stat 500 – Homework 1 (Solutions)

1. There are many ways to do this assignment. Here is one, with commands and some comments.

```
> library(faraway)
> data(teengamb)
# Fix the variable ‘sex’ to be a factor since it is categorical:
> teengamb$sex<-as.factor(teengamb$sex)
> summary(teengamb)
 sex        status          income           verbal          gamble     
 0:28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0  
 1:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1  
        Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0  
        Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3  
        3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4  
        Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0  
```

We notice that there are no missing values. From this summary we see that there were more males sampled than females. Now let’s look at the histograms of the **income** variable and the **gamble** variable.

```
> hist(teengamb$income)
> hist(teengamb$gamble)
```

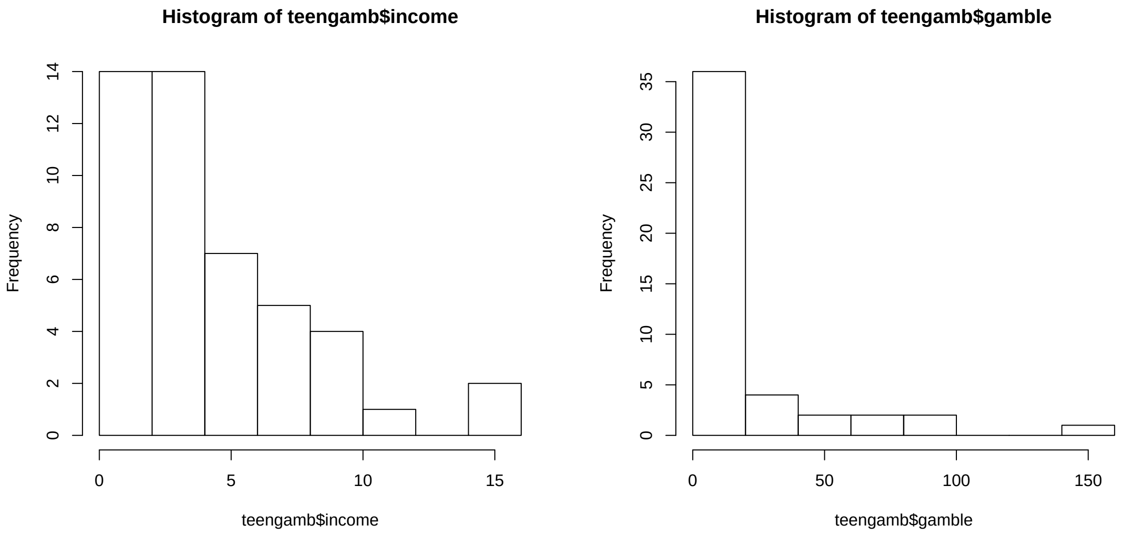
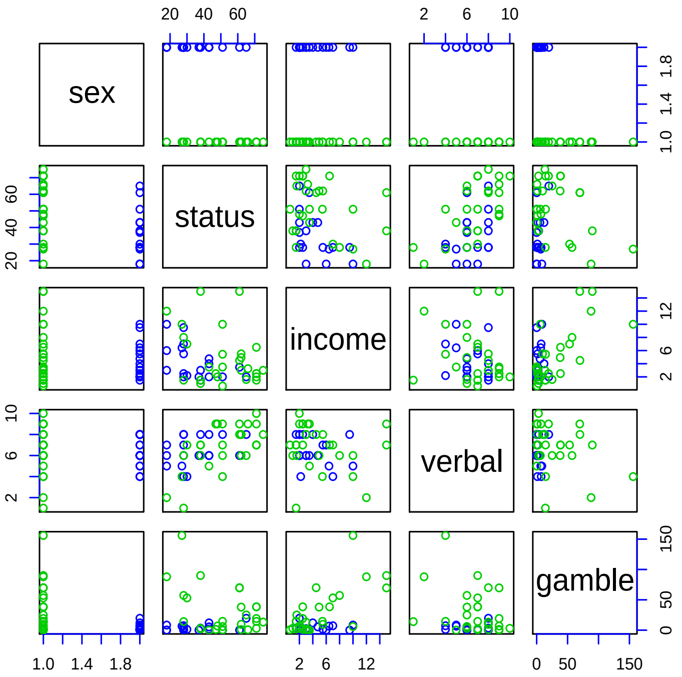


Figure 1: (a) Histogram of **income**, and (b) Histogram of **gamble.**

As the summary above suggested, both the histograms are skewed to the right and there are a few large outliers. Next let’s investigate the pairwise scatter plots of all the variables in this data set, we can color them by the gender.

```
> pairs(teengamb,col=as.numeric(teengamb$sex)+2)
```



It appears in this study that males tend to spend more on gambling than females. Also, the variables **verbal** and **status** look like they may be slightly positively correlated and **gamble** and **income** may also be correlated. The following command confirms those two correlations are greater than 0.5.

```
> cor(teengamb[,-1])
            status     income     verbal      gamble
status  1.00000000 -0.2750340  0.5316102 -0.05042081
income -0.27503402  1.0000000 -0.1755707  0.62207690
verbal  0.53161022 -0.1755707  1.0000000 -0.22005619
gamble -0.05042081  0.6220769 -0.2200562  1.00000000
```

The correlation between **gamble** and **income** makes sense, because people who make more money have more money to spend on gambling. This concludes the preliminary graphical and numerical summary of this data.

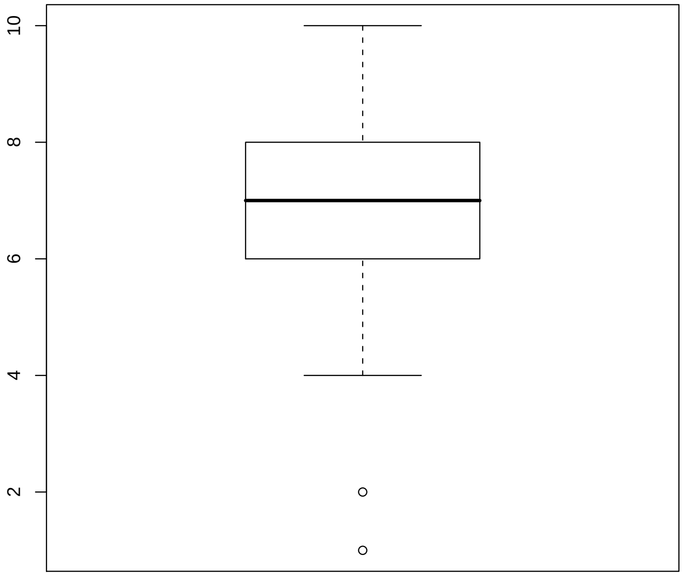
2. We also see that the mean of gamble is much larger than the median, suggesting the distribution is right skewed or may have large outliers, which is likely since the maximum value is so much larger than the other quartile values.

3. There are 9 different values of verbal. They are:

```
> unique(teengamb$verbal)
[1]  8  6  4  7  5  9  2 10  1
> length(unique(teengamb$verbal))
[1] 9
```

4. Consider the following boxplot.

```
> boxplot(teengamb$verbal)
```



From the above boxplot we observe that 1 and 2 are the possible values of outliers.