Stat500(Section002): Homework #1

Due on Spet. 22, 2021 at 11:59pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

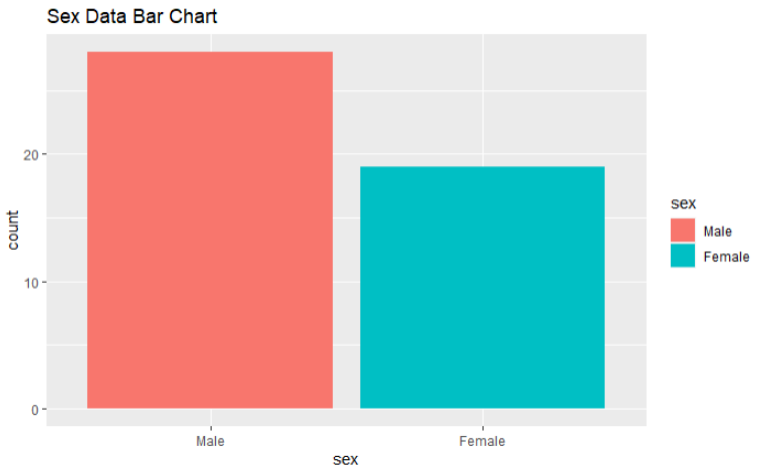
Part 1

We will list some of our codes here, and leave the complete codes on the appendix.We use the following code to transfer sex data, get summary and draw a bar chart for sex data.

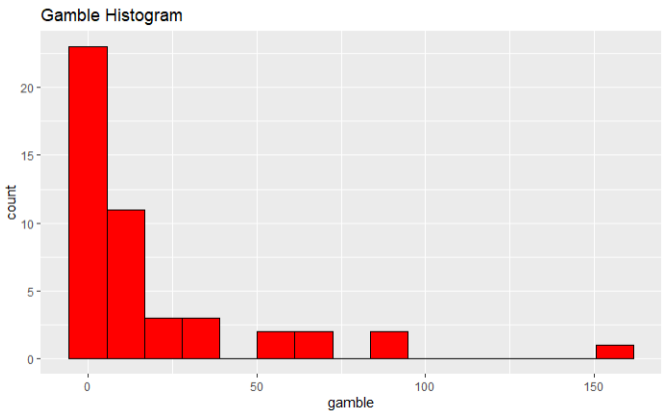
R_code.Rmd

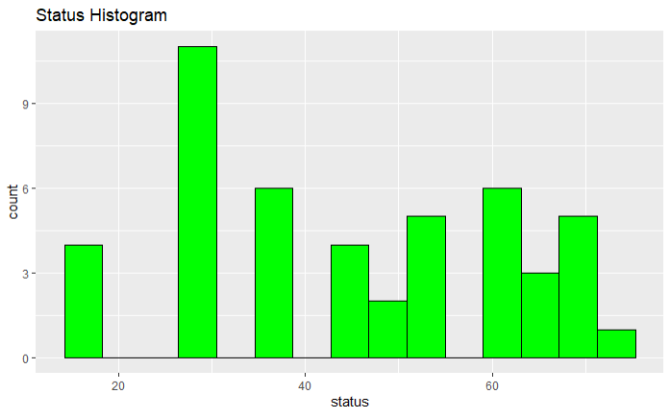
```
teengamb$sex = factor(teengamb$sex)
levels(teengamb$sex) = c('Male','Female')
summary(teengamb)
ggplot(teengamb,aes(x=sex))+geom_bar(aes(fill=sex))
+labs(title = 'Sex Data Bar Chart')
```

```
    sex         status          income          verbal          gamble
Male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
Female: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
```

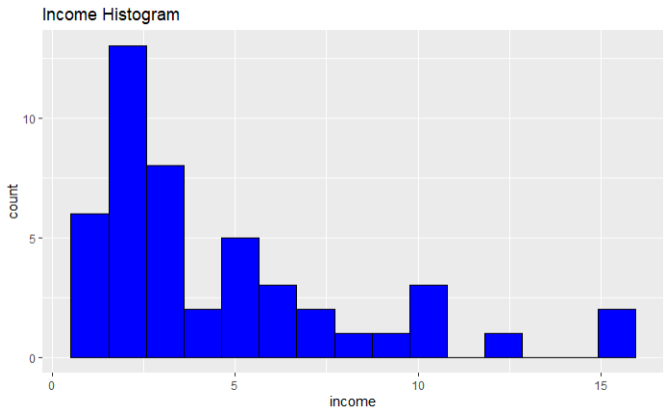


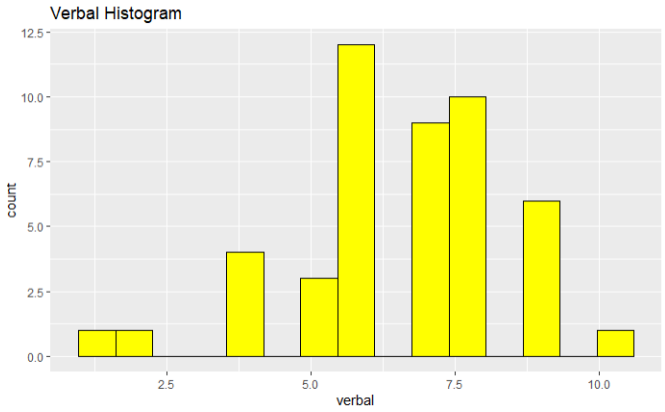
From the bar chart, we can find that the number of male are more than female in this data. Now we draw histogram for the rest data.





From this two histogram,we can see that most people gamble a few, which means it is a little bit skewed. However, there are some outliners. For status, it seems be more uniform and random.

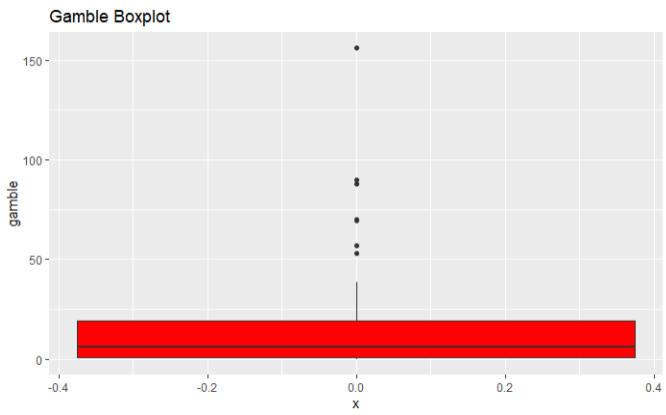


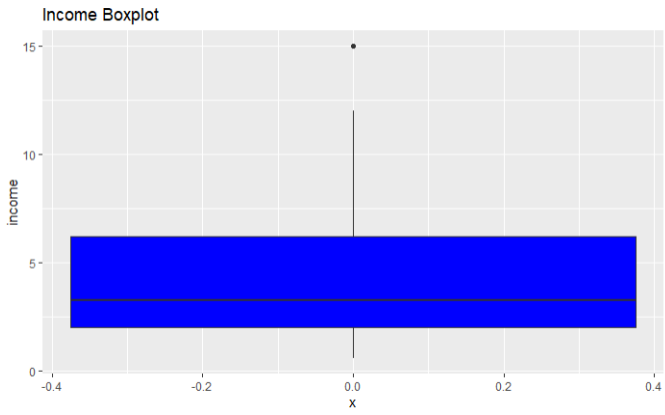


For income, the situation is a more like gamble. Since it is skewed on the left side with some outliners in right side. And the densisy of verbal is more focus on the middle range of values.

Part 2

From the summary we made in the last part,we can know the mean of gamble that is 19.2 is greater than the median of gamble which is 6.0. The mean of income that is 4.64 is also greater than the median which is 3.25. The reason of this situation might be that for income, more poor people want to earn money though gamble, and a few rich people gamble to have fun, there two kind of people lower the median and raise the mean. For gamble, it might be some people gamble more than others a lot to raise the mean. (BTW, that might be the case gamble ruins people I guess).
Now, we draw boxplots for these 2 variables:





we can clearly see that the 50% line in the boxplots are under the intermediate point in the boxplot, Hence I guess the mean is greater than the median, which is the ground truth.

Part 3
we use the following code to get the result:

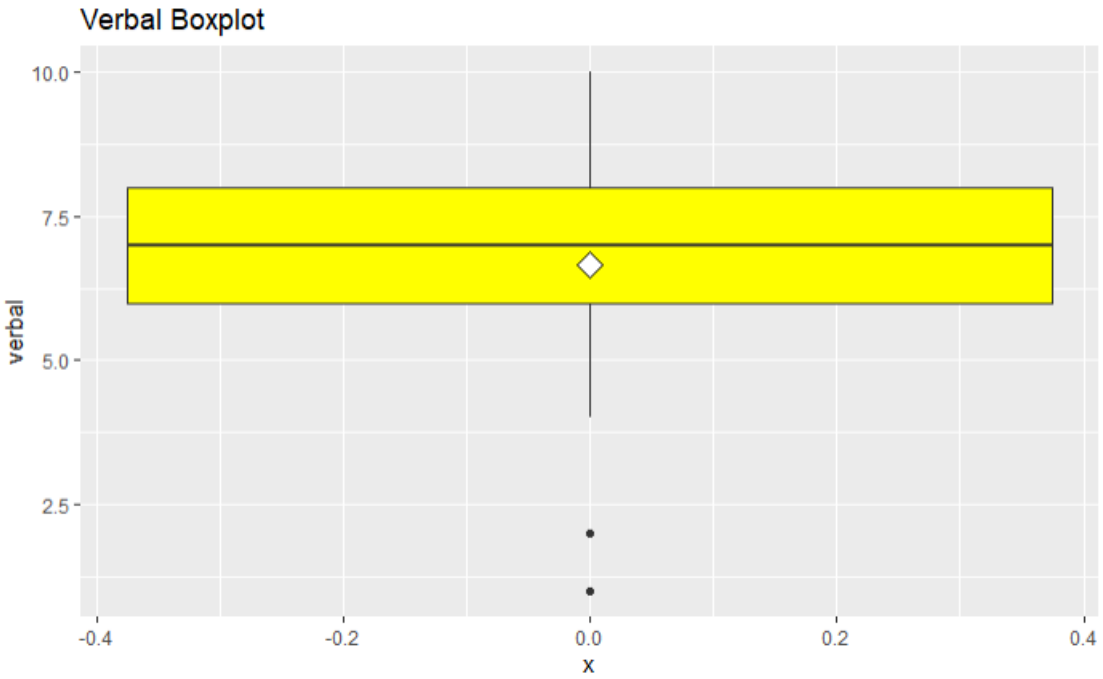
R_code.Rmd

```
uni_ver = unique(teengamb$verbal)
```

```
print(uni_ver)
print(length(uni_ver))
```

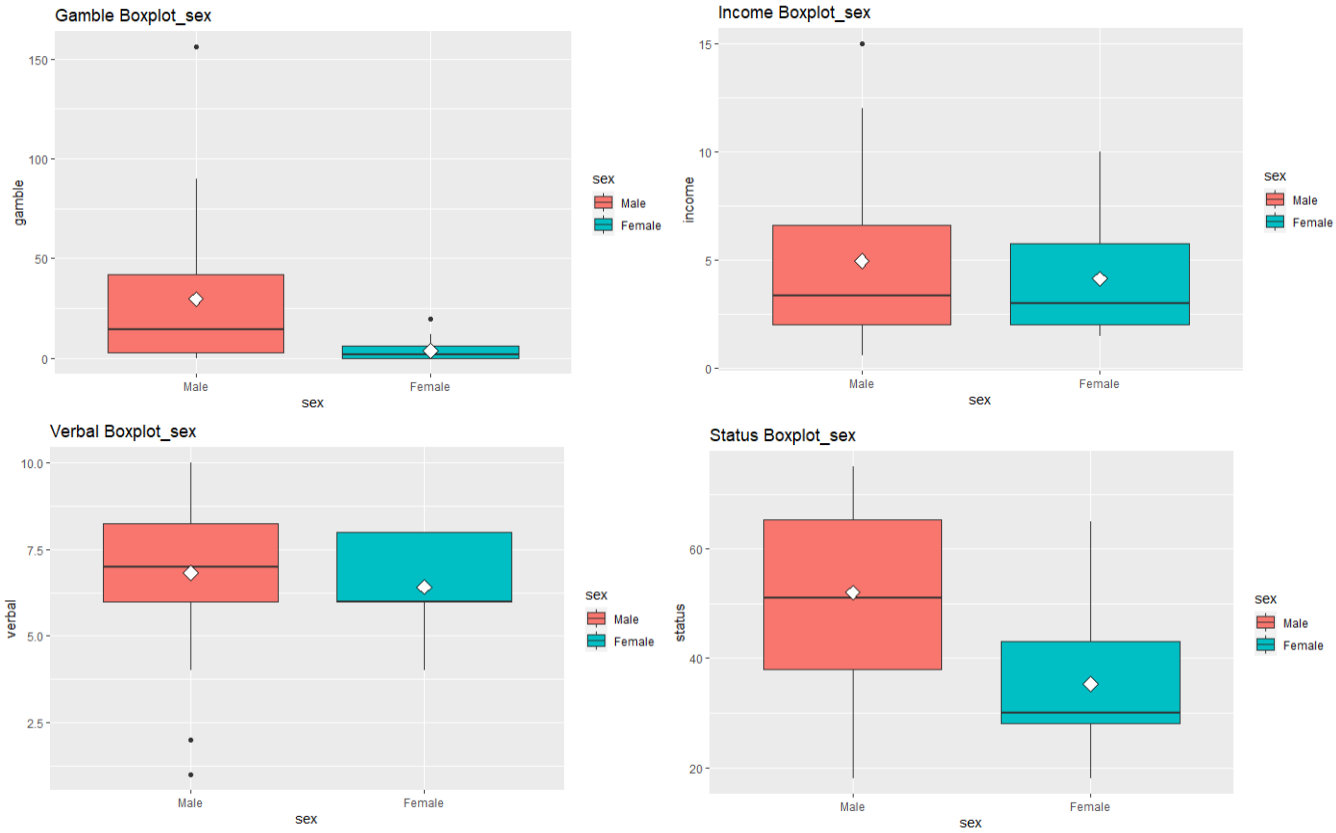
we get there are 9 different values.

Part 4
First, we draw the boxplot of verbal.



where the white diamond is the mean point here. We can see outliner in the boxplot.After inquery the exact value of this two points, we get the possible values of outlying verbal scores is 1 and 2.

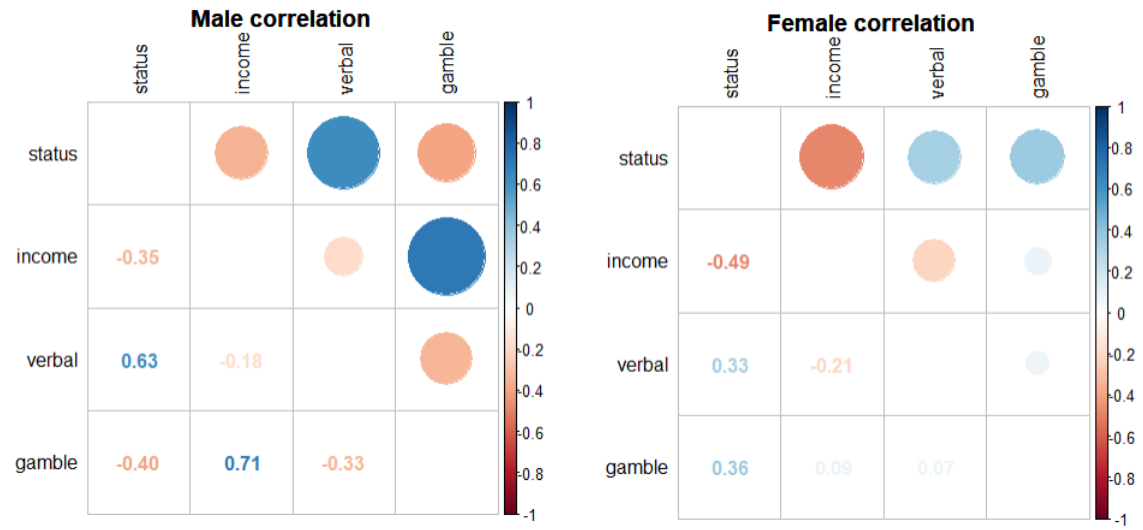
Part 5
Let us draw the boxplot with different sex data.



From the boxplot we can see that:

- Female gamble fewer money than male.Female are more rational than male.
- Female taking the survey have similar income level with male. But they have higher lower bound.
- Female have similar verbal scores level with male. But they have lower upper bound and lower median.
- Male have a higher status than female in total.

Then we draw a figure to show the correlation matrix in different sex. we get:



From this two figures, we can find in male's data, status have high correlation with verbal and income have high correlation with gamble. Andd for female's data, these two correlation is decresing and status have correlation with gamble. Female may think more about their status when they gamble while male may think more about their own income.

appendix

Here is the all code.

R_code.Rmd

```
library(faraway)
library(ggplot2)
library(ggcorrplot)
library(corrplot)
data(teengamb)
var_sex = var(teengamb$sex)
cov_all = cor(teengamb)
corrplot.mixed(cov_all,tl.col = 'black',tl.pos = 'lt')
teengamb$sex = factor(teengamb$sex)
levels(teengamb$sex) = c('Male','Female')
summary(teengamb)
ggplot(teengamb,aes(x=sex))+geom_bar(aes(fill=sex))
+labs(title = 'Sex Data Bar Chart')

ggplot(teengamb, aes(x=gamble)) +
geom_histogram(bins=15, fill='red',col='black') +
labs(title="Gamble Histogram")
ggplot(teengamb, aes(x=status)) +
geom_histogram(bins=15, fill='green',col='black') +
labs(title="Status Histogram")
ggplot(teengamb, aes(x=income)) +
geom_histogram(bins=15, fill='blue',col='black') +
labs(title="Income Histogram")
ggplot(teengamb, aes(x=verbal)) +
geom_histogram(bins=15, fill='yellow',col='black') +
labs(title="Verbal Histogram")

ggplot(teengamb, aes(x=0,y=gamble)) + geom_boxplot(fill='red')+
labs(title="Gamble Boxplot")
ggplot(teengamb, aes(x=0,y=status)) + geom_boxplot(fill='green')+
labs(title="Status Boxplot")
ggplot(teengamb, aes(x=0,y=income)) + geom_boxplot(fill='blue')+
labs(title="Income Boxplot")
ggplot(teengamb, aes(x=0,y=verbal)) + geom_boxplot(fill='yellow')+
labs(title="Verbal Boxplot")

uni_ver = unique(teengamb$verbal)
print(uni_ver)
print(length(uni_ver))
```

```
ggplot(teengamb, aes(x=sex,y=gamble)) + geom_boxplot(aes(fill=sex))+
stat_summary(fun = "mean", geom = "point",shape=23,fill='white',size=4)+
labs(title="Gamble Boxplot_sex")
ggplot(teengamb, aes(x=sex,y=income)) + geom_boxplot(aes(fill=sex))+
stat_summary(fun = "mean", geom = "point",shape=23,fill='white',size=4)+
labs(title="Income Boxplot_sex")
ggplot(teengamb, aes(x=sex,y=verbal)) + geom_boxplot(aes(fill=sex))+
stat_summary(fun = "mean", geom = "point",shape=23,fill='white',size=4)+
labs(title="Verbal Boxplot_sex")
ggplot(teengamb, aes(x=sex,y=status)) + geom_boxplot(aes(fill=sex))+
stat_summary(fun = "mean", geom = "point",shape=23,fill='white',size=4)+
labs(title="Status Boxplot_sex")

male = teengamb[which(teengamb$sex == 'Male'),]
female = teengamb[which(teengamb$sex == 'Female'),]
cor_male = cor(male[,2:5])
cor_female = cor(female[,2:5])
corrplot.mixed(cor_male,tl.col = 'black',tl.pos = 'lt',
title='Male correlation',mar=c(0, 0, 1, 0))
corrplot.mixed(cor_female,tl.col = 'black',tl.pos = 'lt',
title = 'Female correlation',mar=c(0, 0, 1, 0))
```