Testosterone, diversity, and group project performance project

Qiao Su

9/7/2019

Assignment description

See: http://rosmarus.refsmmat.com/datasets/datasets/hormone-diversity/

The questions in the assignment center around the following variables: diversity, composition of groups (Ethnicity, Age, Gender, Country), testosterone: want to predict final.performance. This is The team's final performance score, but unclear if it reliably depends on final.cash, final.rank etc.

Introduction and data summary

```
summary(ind_dat)
```

```
##        ID           team.id         Age           Gender
##  Min.   :102.0   2      :  6   Min.   :23.00   Female:133
##  1st Qu.:343.2   12     :  6   1st Qu.:26.00   Male  :237
##  Median :552.5   24     :  6   Median :27.00
##  Mean   :530.3   35     :  6   Mean   :27.45
##  3rd Qu.:729.8   44     :  6   3rd Qu.:28.00
##  Max.   :874.0   55     :  6   Max.   :37.00
##                  (Other):334   NA's   :13
##             Ethnicity      Cortisol       Testosterone
##  Asian           : 61   Min.   :0.0300   Min.   : 15.28
##  Black           : 17   1st Qu.:0.1060   1st Qu.: 62.58
##  Hispanic        : 40   Median :0.1700   Median :101.24
##  Other           :  9   Mean   :0.2195   Mean   :110.45
##  South Asian     : 35   3rd Qu.:0.2700   3rd Qu.:148.05
##  South East Asian:  5   Max.   :2.1800   Max.   :541.23
##  White           :203   NA's   :5        NA's   :5
##   log.cortisol     log.testosterone      Country
##  Min.   :-3.5066   Min.   :2.727    USA      :213
##  1st Qu.:-2.2443   1st Qu.:4.136    China    : 19
##  Median :-1.7720   Median :4.617    India    : 16
##  Mean   :-1.7627   Mean   :4.534    Korea    : 10
##  3rd Qu.:-1.3093   3rd Qu.:4.998    Argentina:  9
##  Max.   : 0.7793   Max.   :6.294    Canada   :  8
##  NA's   :5         NA's   :5        (Other)  : 95
```

```
summary(team_dat)
```

```
##     team.id     team.size final.performance  time.of.day
##  2      : 1   Min.   :3   Min.   :-3.0807   Min.   : 9.000
##  3      : 1   1st Qu.:5   1st Qu.:-0.4267   1st Qu.: 9.438
##  4      : 1   Median :5   Median : 0.1817   Median :10.750
##  5      : 1   Mean   :5   Mean   : 0.0000   Mean   :11.672
##  6      : 1   3rd Qu.:5   3rd Qu.: 0.6012   3rd Qu.:14.250
##  9      : 1   Max.   :6   Max.   : 1.1099   Max.   :16.000
```

```
##  (Other):68
##     females        final.cash      final.contracts final.reorders
##  Min.   :0.000   Min.   : 642783   Min.   :1.000   Min.   : 15.00
##  1st Qu.:2.000   1st Qu.:1362974   1st Qu.:2.000   1st Qu.: 81.25
##  Median :2.000   Median :1664432   Median :3.000   Median : 86.00
##  Mean   :1.784   Mean   :1600262   Mean   :2.662   Mean   : 84.54
##  3rd Qu.:2.000   3rd Qu.:1820144   3rd Qu.:3.000   3rd Qu.: 90.00
##  Max.   :2.000   Max.   :2050636   Max.   :3.000   Max.   :110.00
##
##    final.rank     interim.performance  interim.cash     interim.contracts
##  Min.   : 1.000   Min.   :-2.1978     Min.   : 396109   Min.   :1.000
##  1st Qu.: 4.000   1st Qu.:-0.2651     1st Qu.: 734886   1st Qu.:2.000
##  Median : 7.500   Median : 0.1456     Median : 806530   Median :3.000
##  Mean   : 7.257   Mean   : 0.0000     Mean   : 812429   Mean   :2.404
##  3rd Qu.:10.000   3rd Qu.: 0.6604     3rd Qu.: 925021   3rd Qu.:3.000
##  Max.   :14.000   Max.   : 1.0924     Max.   :1062138   Max.   :3.000
##                   NA's   :22          NA's   :22        NA's   :22
##  interim.reorders  interim.rank
##  Min.   : 20.00   Min.   : 1.00
##  1st Qu.: 75.75   1st Qu.: 4.00
##  Median : 85.00   Median : 8.00
##  Mean   : 81.40   Mean   : 8.00
##  3rd Qu.: 90.00   3rd Qu.:11.25
##  Max.   :108.00   Max.   :15.00
##  NA's   :22       NA's   :22
```

```
# seems like interim.* columns contain a lot of missing data.
```

Questions

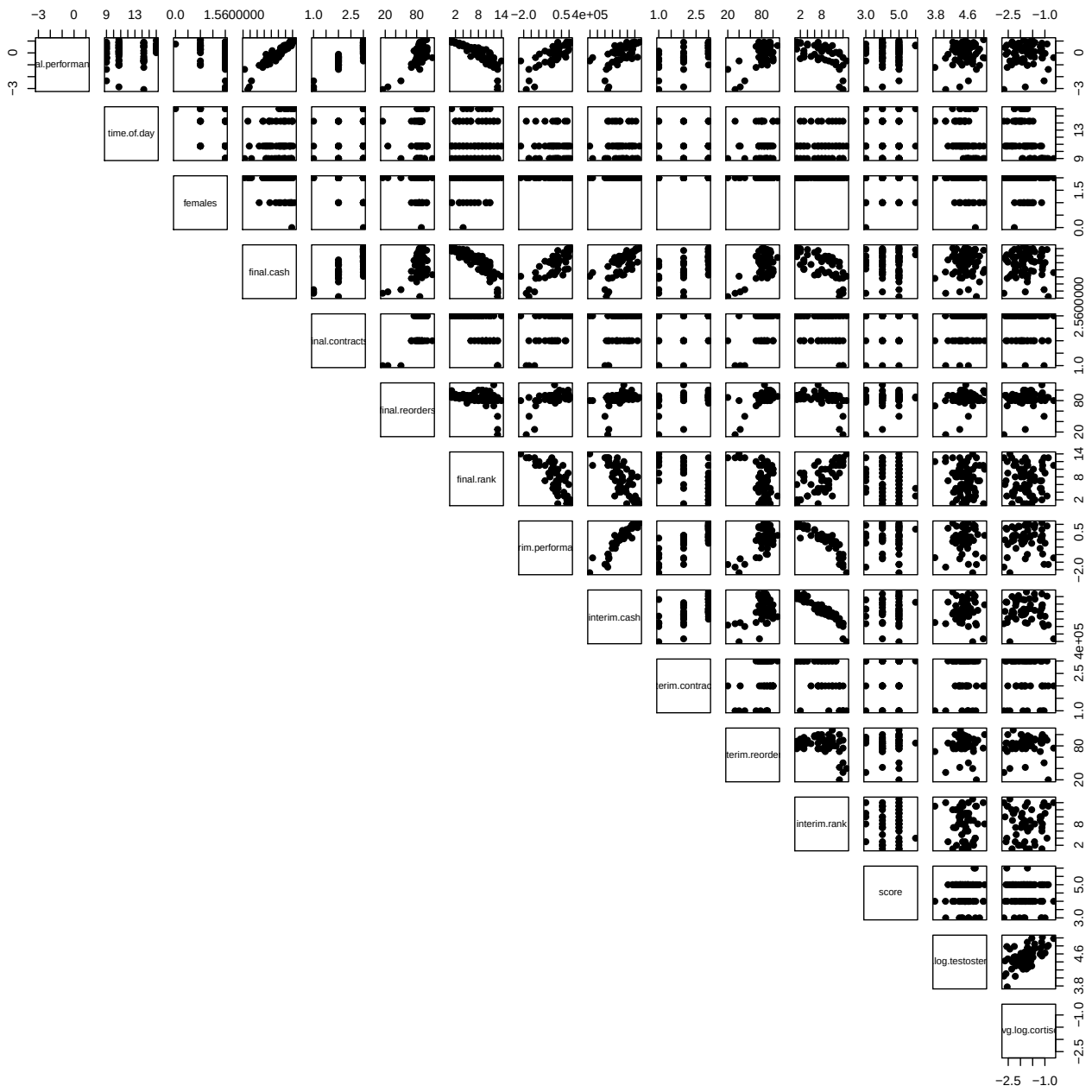
Q1

The original paper measured the diversity of each group using something called "group faultline analysis", which looks at the group members' genders, countries of origin, and ethnicities to calculate a diversity score. The calculation is somewhat involved, so we'll make a simpler score.

For each group, calculate the number of unique gender-ethnicity-country combinations (such as female-white-Russia or male-Indian-USA) among the group members, and store this with the other group information such as team size and performance. Also calculate the average testosterone level for each group.

```
get_score <-function(group){
  score <- length(unique(ind_dat$combo[ind_dat$team.id == group,]))
  return(score)
}
#  calculate the number of unique gender-ethnicity-country combinations
ind_dat$combo <-paste(ind_dat$Gender, ind_dat$Ethnicity, ind_dat$Country)
team_dat$score<- unlist(lapply(team_dat$team.id,
               function(x){length(unique(ind_dat$combo[ind_dat$team.id == x]))}))

#  calculate the average testosterone level for each group.
team_dat$avg.log.testosterone<- unlist(lapply(team_dat$team.id,
               function(x){mean(ind_dat$log.testosterone[ind_dat$team.id == x], na.rm = TRUE)}))

#  calculate the average cortisol level for each group.
team_dat$avg.log.cortisol<- unlist(lapply(team_dat$team.id,
               function(x){mean(ind_dat$log.cortisol[ind_dat$team.id == x], na.rm = TRUE)}))
```

```
# check relationships of all variables of interest
vars <- colnames(team_dat)[c(3:17)]
#cor(team_dat)
pairs(team_dat[vars], pch = 19,  lower.panel=NULL)
```

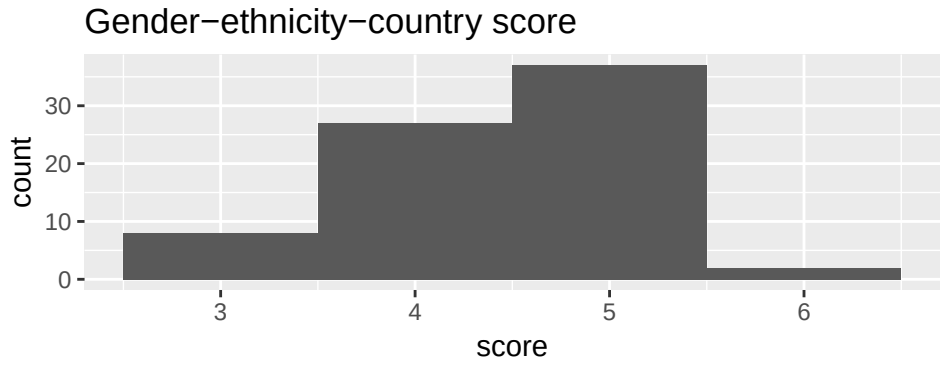


Q2

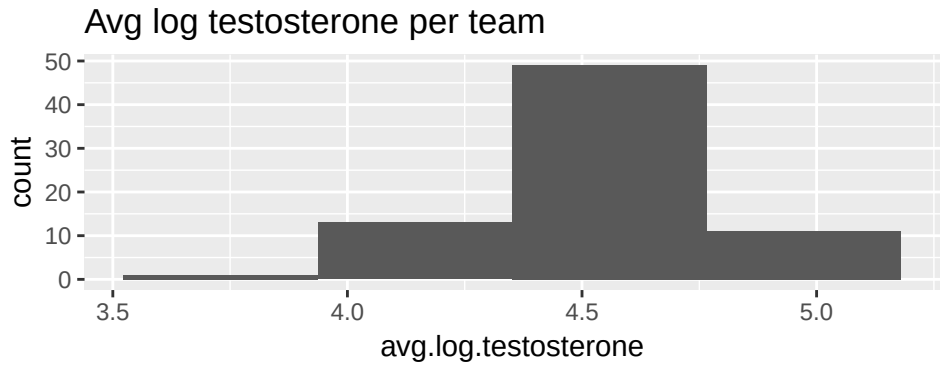
Do exploratory data analysis to explore the composition of groups, the typical amount of diversity, and the typical amounts of testosterone. Note particularly that the data includes the logs of the cortisol and testosterone levels as well as the raw levels; does your EDA suggest you should use the logs or the raw values?

Composition of groups

```
# visualise the distribution of diversity
ggplot(team_dat, aes(x = score)) +
  geom_histogram(bins = 4)+ labs(title="Gender-ethnicity-country score")
```

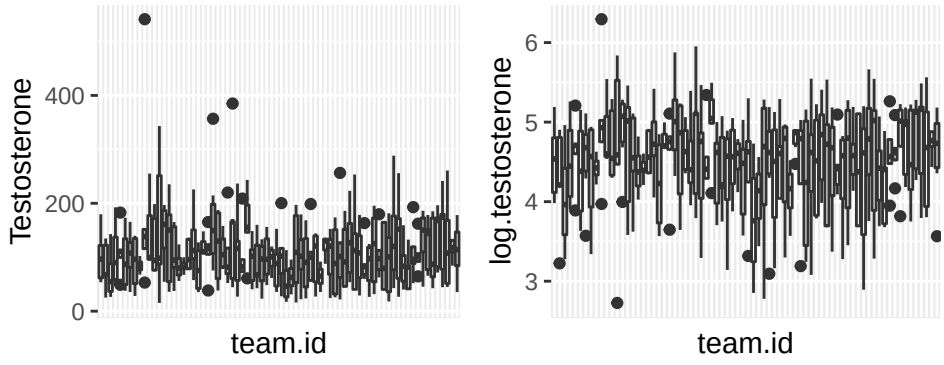



```
# visualise the distribution of testosterone
ggplot(team_dat, aes(x = avg.log.testosterone)) +
  geom_histogram(bins = 4)+ labs(title="Avg log testosterone per team")
```

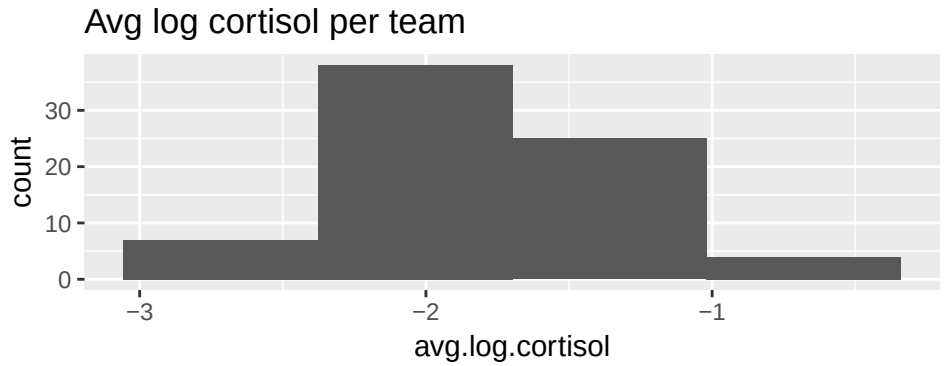



```
p1 <-ggplot(ind_dat, aes(x= team.id, y=Testosterone))+
  geom_boxplot()+theme_hw

p2 <-ggplot(ind_dat, aes(x= team.id, y=log.testosterone))+
  geom_boxplot()+theme_hw

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).

## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

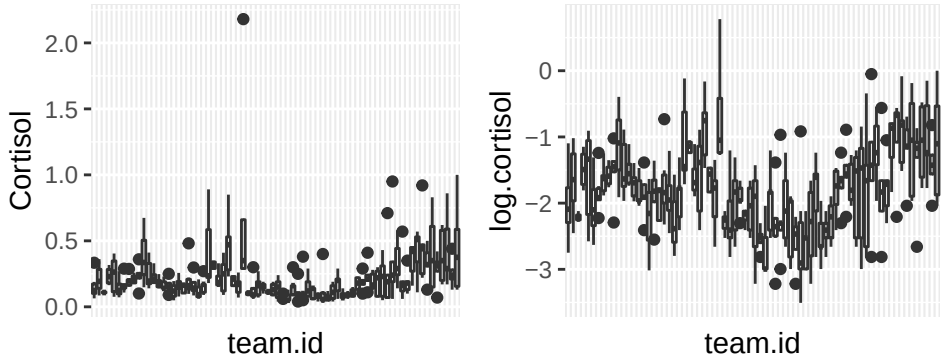


```r
# visualise the distribution of cortisol
ggplot(team_dat, aes(x = avg.log.cortisol)) +
  geom_histogram(bins = 4)+ labs(title="Avg log cortisol per team")
```




```r
p1 <-ggplot(ind_dat, aes(x= team.id, y=Cortisol))+
  geom_boxplot()+theme_hw

p2 <-ggplot(ind_dat, aes(x= team.id, y=log.cortisol))+
  geom_boxplot()+theme_hw

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).

## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



What's up with the missing testosterone data?

```r
#head(ind_dat[!complete.cases(ind_dat$log.testosterone),])
head(team_dat[!complete.cases(team_dat),])
```

```
##   team.id team.size final.performance time.of.day females final.cash
## 1       2         6         0.5055488       14.25       2    1805166
## 2       3         5         0.6040071       14.25       1    1894642
## 3       4         4        -0.4507569       14.25       2    1338256
## 4       5         4         0.9101548       14.25       2    1998412
## 5       6         5         0.8315757       14.25       1    1983439
## 6       9         4        -0.6797941       14.25       1    1524141
##   final.contracts final.reorders final.rank interim.performance
## 1               3             87          5                  NA
## 2               3             85          4                  NA
## 3               3             80         12                  NA
```

```
## 4               3            90          2                  NA
## 5               3            90          3                  NA
## 6               2            77         10                  NA
##   interim.cash interim.contracts interim.reorders interim.rank score
## 1           NA                NA               NA           NA     4
## 2           NA                NA               NA           NA     4
## 3           NA                NA               NA           NA     4
## 4           NA                NA               NA           NA     4
## 5           NA                NA               NA           NA     4
## 6           NA                NA               NA           NA     3
##   avg.log.testosterone avg.log.cortisol
## 1             4.519117       -2.008135
## 2             4.268671       -1.667746
## 3             4.041234       -2.211397
## 4             4.515612       -1.485606
## 5             4.624405       -1.609215
## 6             4.387938       -2.035934
```

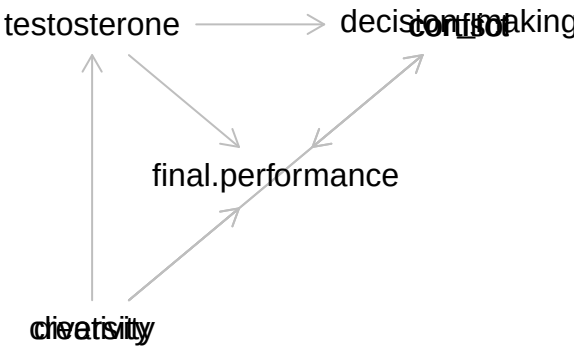
Sketch out causal diagrams

Some research has suggested that groups of people working on a task can do better if the group is more diverse, since diverse group members can suggest more creative ideas and make better decisions. At the same time, diverse groups can have more conflict than less diverse groups, possibly eliminating those benefits.

People with high levels of testosterone tend to be competitive and may try to dominate their groups, leading to conflict, while people with low testosterone may be more cooperative. So do groups with low testosterone work better together, and benefit more from diversity, than those with high testosterone?

The findings suggest that diversity is beneficial for performance, but only if group-level testosterone is low; diversity has a negative effect on performance if group-level testosterone is high.

```
library(dagitty)

g <- dagitty('dag {
    testosterone  [pos="0,0"]
    diversity [pos="0,1"]
    final.performance [pos="1,0.5"]
    cortisol [pos="2,0"]
    conflict[pos="2,0"]
    decision_making[pos="2,0"]
    creativity[pos="0,1"]

    diversity-> decision_making
    decision_making -> final.performance<- creativity
    diversity -> testosterone -> decision_making -> final.performance
    diversity -> cortisol-> final.performance
    testosterone -> final.performance <- diversity
    cortisol->final.performance
}')
plot(g)
```



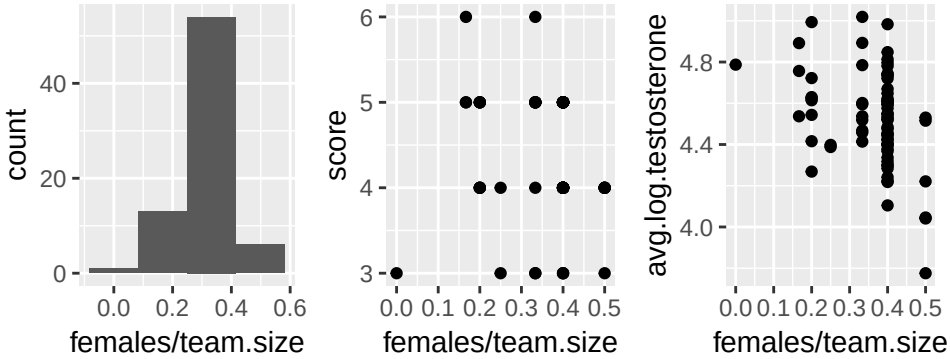

```
impliedConditionalIndependencies( g )
```

```
## conflict _||_ cortisol
## conflict _||_ creativity
## conflict _||_ decision_making
## conflict _||_ diversity
## conflict _||_ final.performance
## conflict _||_ testosterone
## cortisol _||_ creativity
## cortisol _||_ decision_making | diversity
## cortisol _||_ testosterone | diversity
## creativity _||_ decision_making
## creativity _||_ diversity
## creativity _||_ testosterone
```

Q3

The data also includes participant ages. It's possible age is related to hormone levels, as is gender, and that both are related to final performance (perhaps older team members have more relevant experience, for example). Consider carefully whether you want to include these variables in the model, and how you should include them — average team member age? Gender proportion? #### It's possible age is related to hormone levels? performance?

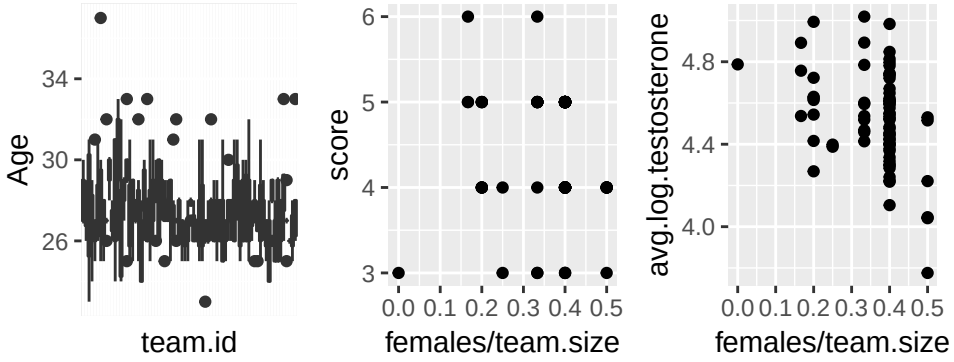
```
p1 <-ggplot(team_dat, aes(x = females/team.size)) +
  geom_histogram(bins =4)
p2 <-ggplot(team_dat, aes(x = females/team.size, y = score)) +
  geom_point()
p3 <-ggplot(team_dat, aes(x = females/team.size, y = avg.log.testosterone)) +
  geom_point()
grid.arrange(p1, p2, p3, ncol = 3)
```



age related to hormone levels? performance?

```
p1 <-ggplot(ind_dat, aes(x= team.id, y=Age))+
  geom_boxplot()+theme_hw
p2 <-ggplot(team_dat, aes(x = females/team.size, y = score)) +
  geom_point()
p3 <-ggplot(team_dat, aes(x = females/team.size, y = avg.log.testosterone)) +
  geom_point()
grid.arrange(p1, p2, p3, ncol = 3)
```

```
## Warning: Removed 13 rows containing non-finite values (stat_boxplot).
```



Q4

Build a model predicting group performance (final.performance) using the group's diversity score (be sure to control for the size of the group) and its average testosterone level. Is there an interaction between the two? Do your results resemble those presented by the original study?

Q5