EDA

Anni Hong

9/15/2020

Basic Info:

```
## [1] 370  10
```

```
## [1] 74 14
```

```
##        ID           team.id            Age          Gender         
##  Min.   :102.0   Min.   :  2.00   Min.   :23.00   Length:370        
##  1st Qu.:343.2   1st Qu.: 35.00   1st Qu.:26.00   Class :character  
##  Median :552.5   Median : 63.00   Median :27.00   Mode  :character  
##  Mean   :530.3   Mean   : 60.79   Mean   :27.45                     
##  3rd Qu.:729.8   3rd Qu.: 87.75   3rd Qu.:28.00                     
##  Max.   :874.0   Max.   :111.00   Max.   :37.00                     
##                                   NA's   :13                        
##   Ethnicity           Cortisol       Testosterone     log.cortisol    
##  Length:370         Min.   :0.0300   Min.   : 15.28   Min.   :-3.5066  
##  Class :character   1st Qu.:0.1060   1st Qu.: 62.58   1st Qu.:-2.2443  
##  Mode  :character   Median :0.1700   Median :101.24   Median :-1.7720  
##                     Mean   :0.2195   Mean   :110.45   Mean   :-1.7627  
##                     3rd Qu.:0.2700   3rd Qu.:148.05   3rd Qu.:-1.3093  
##                     Max.   :2.1800   Max.   :541.23   Max.   : 0.7793  
##                     NA's   :5        NA's   :5        NA's   :5        
##  log.testosterone   Country         
##  Min.   :2.727    Length:370        
##  1st Qu.:4.136    Class :character  
##  Median :4.617    Mode  :character  
##  Mean   :4.534                      
##  3rd Qu.:4.998                      
##  Max.   :6.294                      
##  NA's   :5
```

```
##     team.id         team.size final.performance  time.of.day    
##  Min.   :  2.00   Min.   :3   Min.   :-3.0807   Min.   : 9.000  
##  1st Qu.: 34.25   1st Qu.:5   1st Qu.:-0.4267   1st Qu.: 9.438  
##  Median : 62.50   Median :5   Median : 0.1817   Median :10.750  
##  Mean   : 60.08   Mean   :5   Mean   : 0.0000   Mean   :11.672  
##  3rd Qu.: 86.75   3rd Qu.:5   3rd Qu.: 0.6012   3rd Qu.:14.250  
##  Max.   :111.00   Max.   :6   Max.   : 1.1099   Max.   :16.000  
##                                                                 
##     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
```

Individual dataset

- The individual dataset contains 370 observations, and 10 variables :
  ID Participant ID number team.id ID number of the team this participant belonged to Age Age, in years Gender Gender (Male or Female) Ethnicity Ethnicity of the participant Cortisol Participant's cortisol levels, nMol/L Testosterone Participant's testosterone levels, pg/mL log.cortisol Natural logarithm of the participant's cortisol level log.testosterone Natural logarithm of the participant's testosterone level Country Country of citizenship of the participant
- 18 rows contain at least one missing value in one of the columns

Team dataset

- The team dataset contains 74 teams and 14 variables:
  team.id Team ID number team.size Number of people on the team final.performance The team's final performance score time.of.day The time of day the team's hormone sample was collected (hh.mm) females Number of females in the group final.cash Total cash earned by the team final.contracts Total number of contracts won by the team final.reorders Total number of reorders won by the team final.rank Team's final rank at the end of the project, relative to other teams in their class section interim.performance Same as above, but measured at Day 5 of the study (missing for some teams)
  interim.cash
  interim.contracts
  interim.reorders
  interim.rank
- 22 teams have no interim.rank, interim.reorders, interim.contracts, interim.cash,and interim.performance

```r
#factorize certain variables and create diversity score for each group
indi_dat <- indi_dat %>%
  mutate_if(sapply(indi_dat, is.character), as.factor) %>%
  group_by(team.id) %>%
  mutate(diversity_score = n_distinct(Gender, Ethnicity, Country))
```

```
#intermidiate step of calculating aggregated statistics by group
agg_indi_dat <- indi_dat %>%
  group_by(team.id) %>%
  summarise(mean_testo = mean(Testosterone), mean_log_testo = mean(log.testosterone), sd_testo = sd(Tes
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

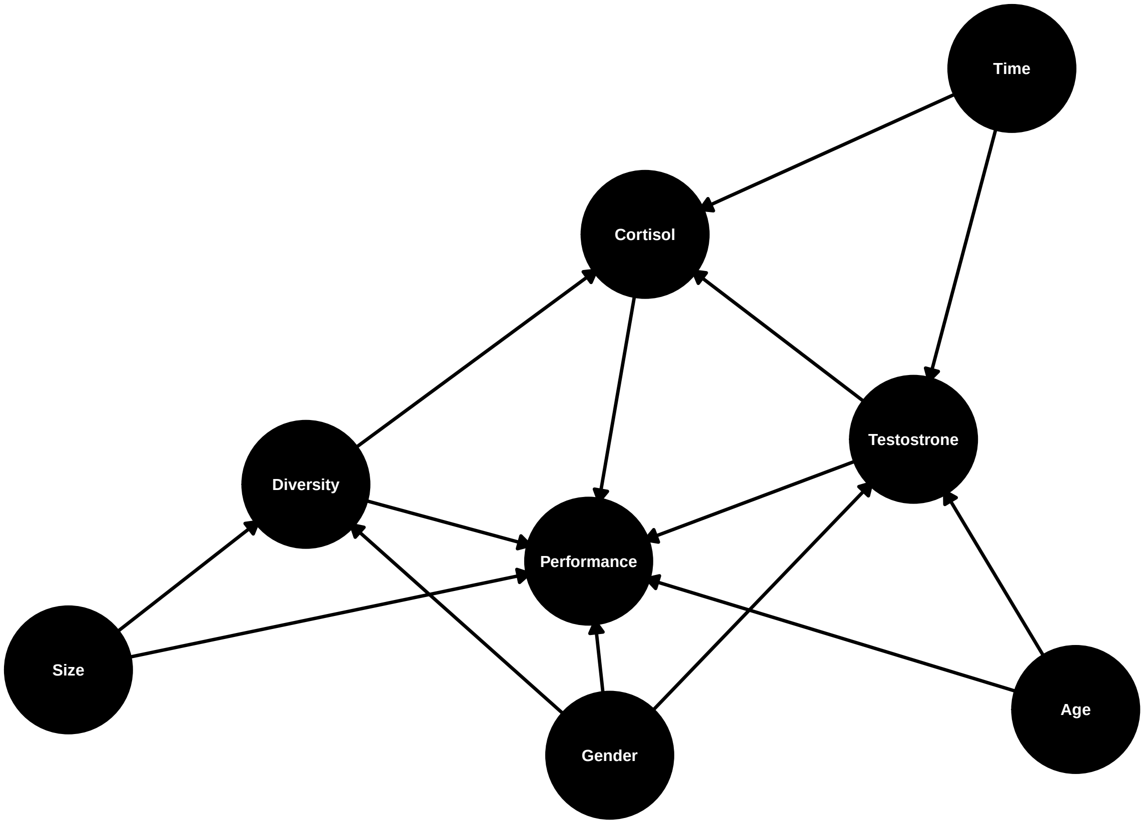
```
#combing the info from the individual dataset to the team dataset, average/sd cortisol and testosterone,
combo_dat <- team_dat %>%
  inner_join(agg_indi_dat, by = "team.id") %>%
  mutate(proportion_female = females/team.size)

write_csv(combo_dat, "./data/combined_processed.csv")
```

Causal DAG

```
tidy_ggdag <- dagify(
  Testostrone ~ Time + Gender + Age,
  Cortisol ~ Time + Testostrone + Diversity,
  Diversity ~ Size + Gender,
  Performance ~ Testostrone + Cortisol + Diversity + Size + Age + Gender,
  exposure = "Diversity",
  outcome = "Performance"
) %>%
  tidy_dagitty()

ggdag(tidy_ggdag, node_size = 22, text_size = 2.2) +
  theme_dag()
```



Understanding the variables

Typical amount of diversity present:

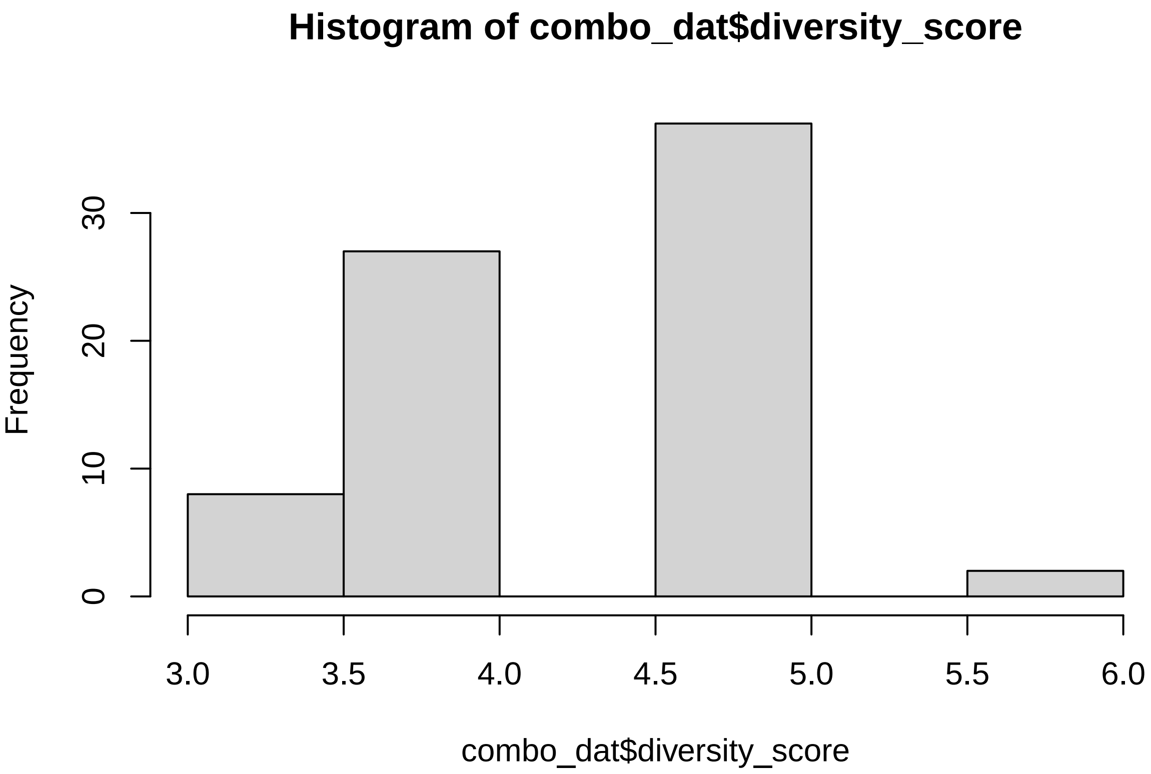
```
diversity_vars <- indi_dat %>% select(Gender, Ethnicity, Country)
```

```
## Adding missing grouping variables: `team.id`
```

```
summary(diversity_vars[,-1])
```

```
##     Gender                 Ethnicity        Country   
##  Female:133   Asian           : 61   USA      :213  
##  Male  :237   Black           : 17   China    : 19  
##               Hispanic        : 40   India    : 16  
##               Other           :  9   Korea    : 10  
##               South Asian     : 35   Argentina:  9  
##               South East Asian:  5   Canada   :  8  
##               White           :203   (Other)  : 95
```

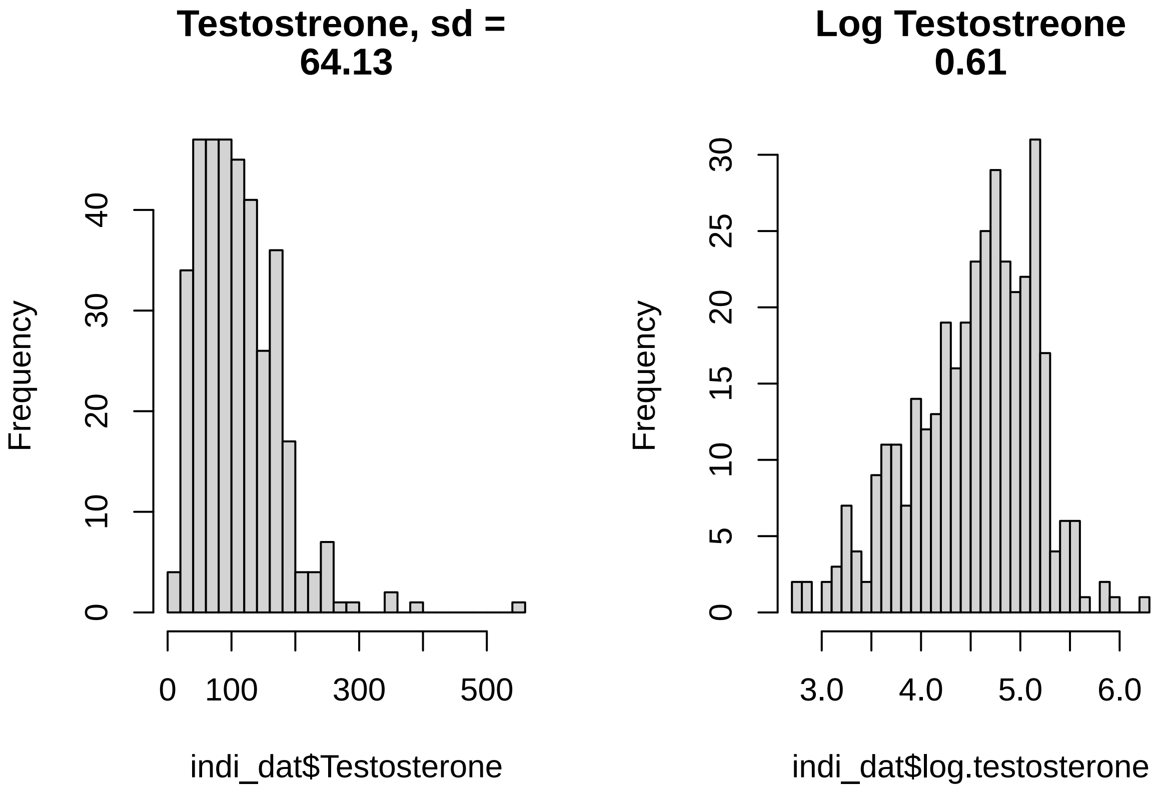
```
hist(combo_dat$diversity_score)
```



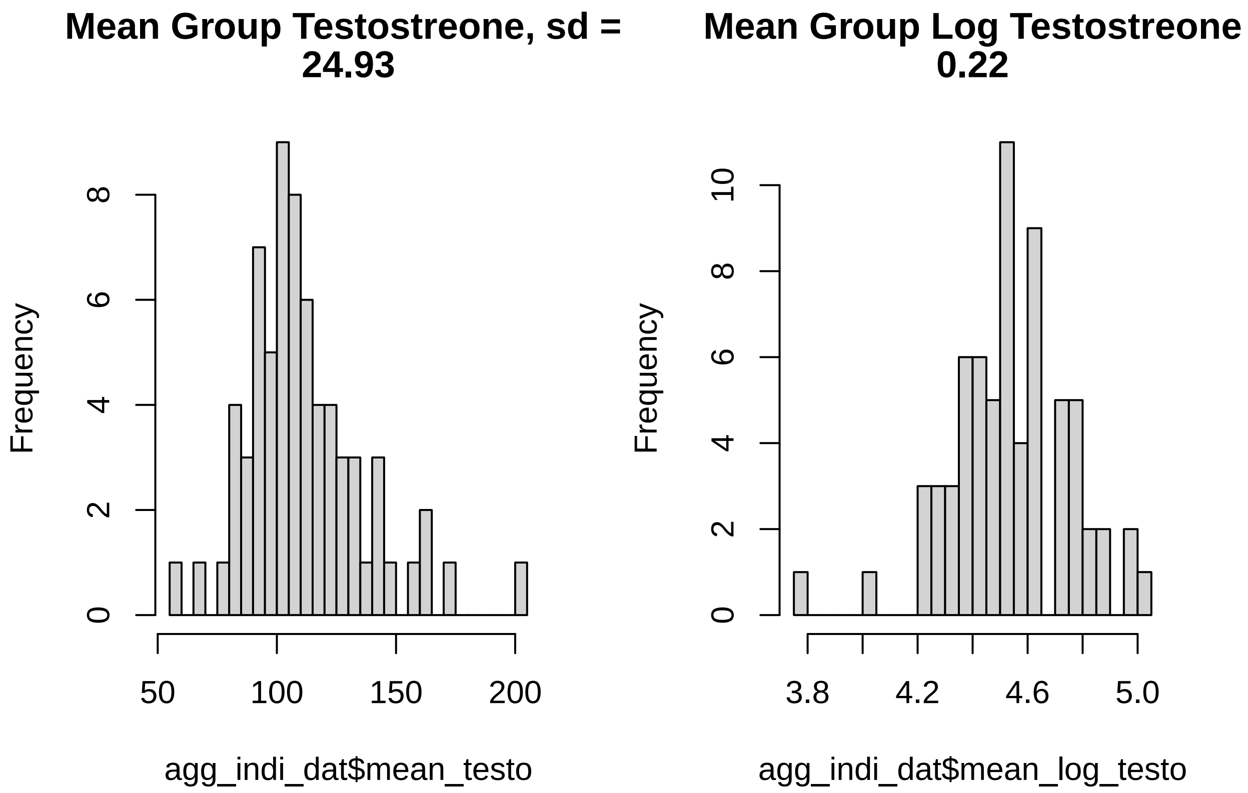
This dataset has a lot more men than women and mostly white Americans. There is a adequate amount of variability in the diversity scores.

testostreone vs log testostrone

```
par(mfrow=c(1,2))
sd_testo = round(sd(indi_dat$Testosterone, na.rm = T),2)
sd_log_testo = round(sd(indi_dat$log.testosterone, na.rm = T),2)
hist(indi_dat$Testosterone, breaks = 30, main = c("Testostreone, sd = ",sd_testo))
hist(indi_dat$log.testosterone, breaks = 30, main = c("Log Testostreone", sd_log_testo))
```



```
par(mfrow=c(1,2))
sd_testo = round(sd(agg_indi_dat$mean_testo, na.rm = T),2)
sd_log_testo = round(sd(agg_indi_dat$mean_log_testo, na.rm = T),2)
hist(agg_indi_dat$mean_testo, breaks = 30, main = c("Mean Group Testostreone, sd = ",sd_testo))
hist(agg_indi_dat$mean_log_testo, breaks = 30, main = c("Mean Group Log Testostreone", sd_log_testo))
```



The above plots shows the histogram of the testostrone level compared to the log of the testostrone level, for overall as well as group mean. The log transformation helps with lessen the impact of outliers on our analysis. We don't want to let the group that contain the individual with really high testostrone level impact our analysis disproportionally since mean is very sensitive to outliers. The standard deviation is drastically reduced in the log transformed variable in both cases and the histogram approximate a normal distribution.

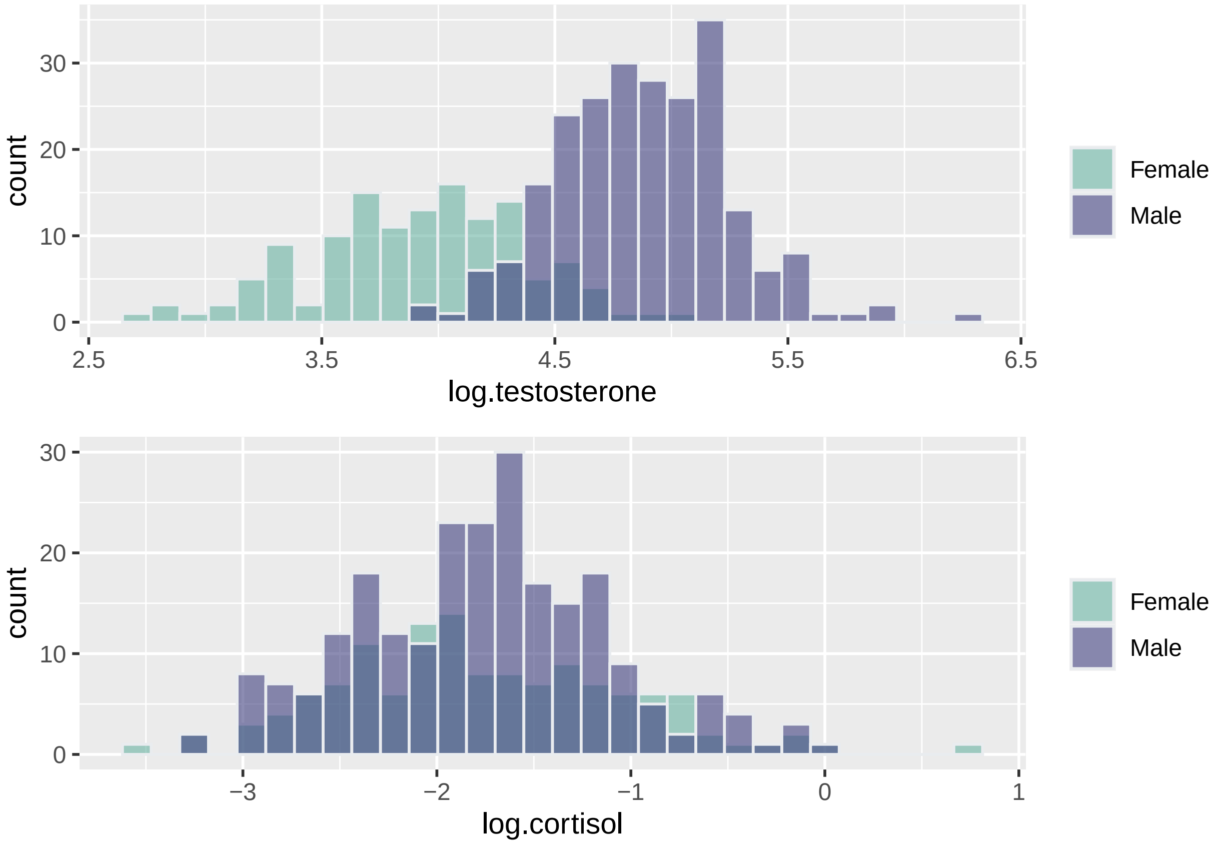
testostrone, cortisol by gender

```r
testo <- indi_dat %>%
  ggplot( aes(x=log.testosterone, fill=Gender)) +
    geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="")

cortisol <- indi_dat %>%
  ggplot( aes(x=log.cortisol, fill=Gender)) +
    geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="")

grid.arrange(testo, cortisol)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

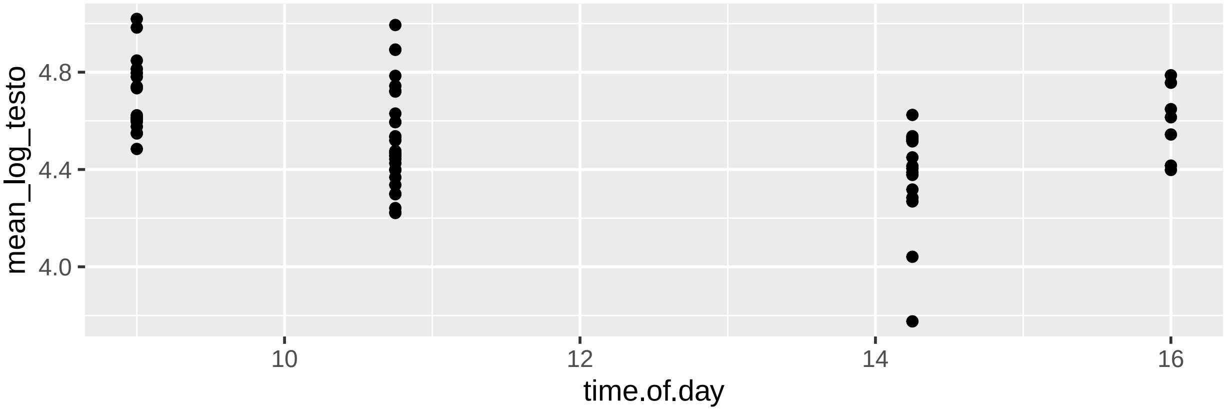


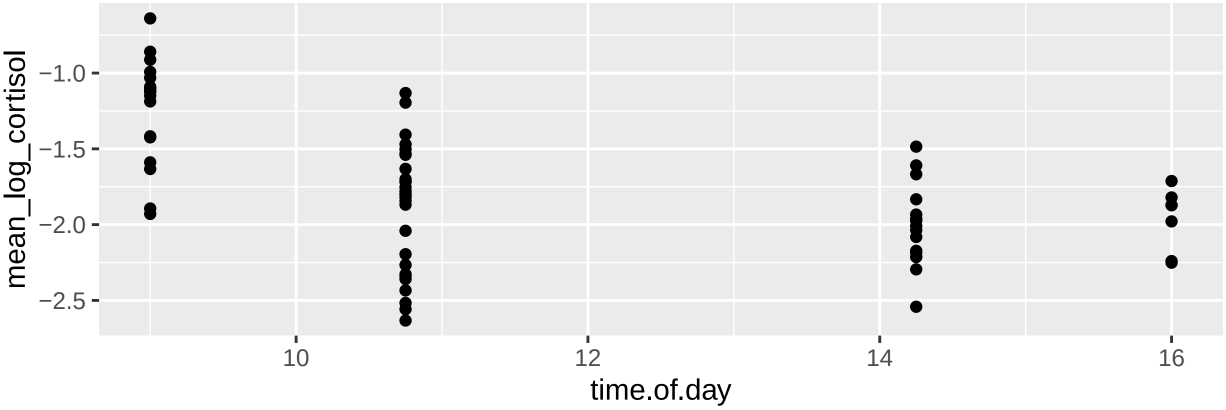
It makes sense that gender is correlationed to the level of testostrone so special attention, since women tend to have lower testostrone level than men. Gender seem to have no impact on crotisol level. ### testostrone, cortisol by time of the day

```
testo <- combo_dat %>%
  ggplot(aes(x=time.of.day, y=mean_log_testo)) +
    geom_point()

cortisol <- combo_dat %>%
  ggplot(aes(x=time.of.day, y=mean_log_cortisol)) +
    geom_point()

grid.arrange(testo, cortisol)
```

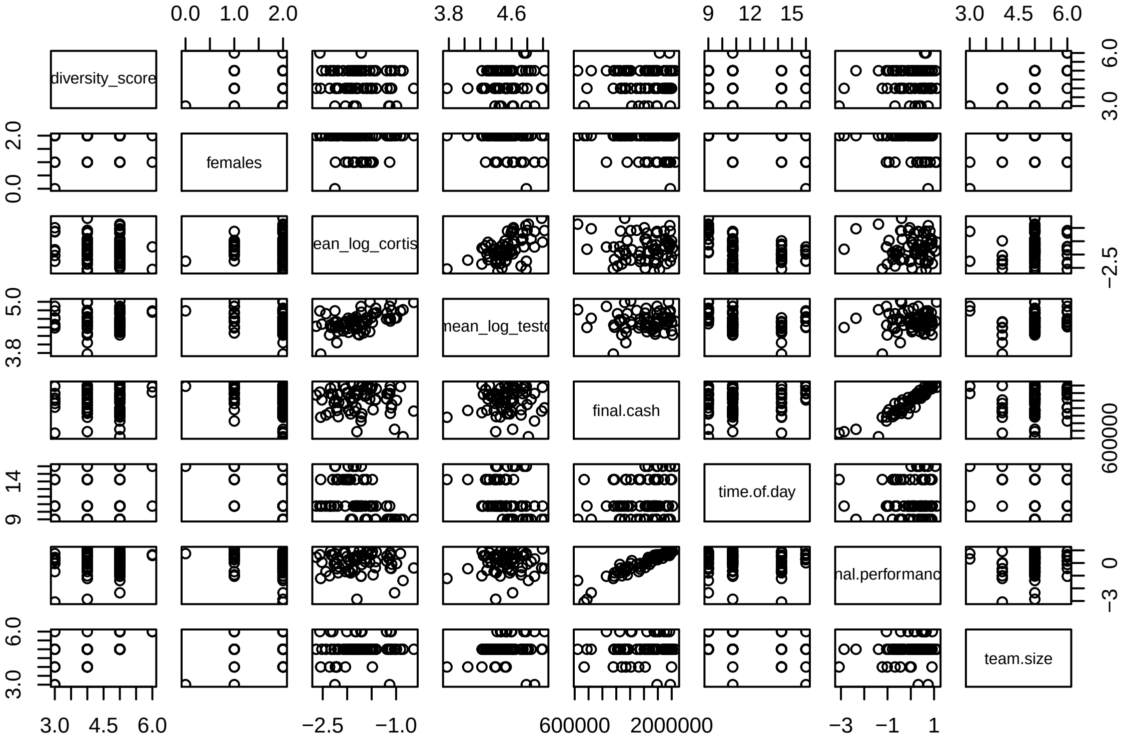




Time of the day could also be a confounder since the testostrone and cortisol level could change base on what time of the day you measure it.

relationships between a subset of variables

```
combo_dat_subset <- combo_dat %>% select(diversity_score,females, mean_log_cortisol, mean_log_testo,  fi
pairs(combo_dat_subset) # not including team id
```



- Worth noting that there are exactly 2 females in all the groups, if we want to control for gender it would make more sense to use percentage.

- Cortisol seems to be correlated with Testostrone levels, since Cortisol is related to stress maybe it will help test/explain why there could be an interaction effect between diversity and testostrone (through causing stress on the team).