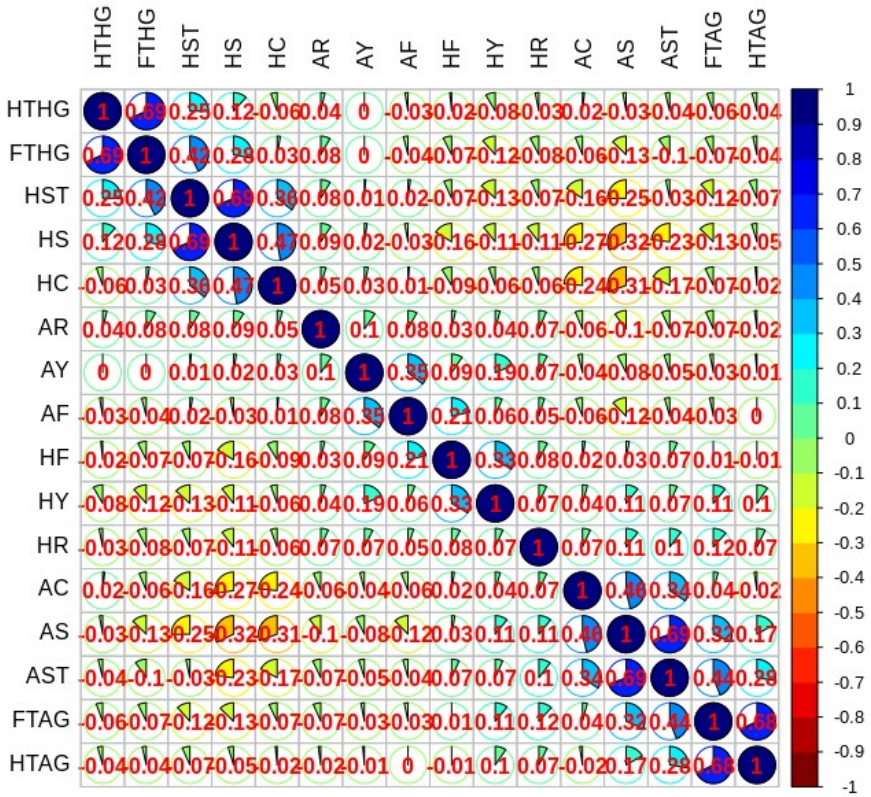
Ярылгасимов Султан
Домашнее задание №8

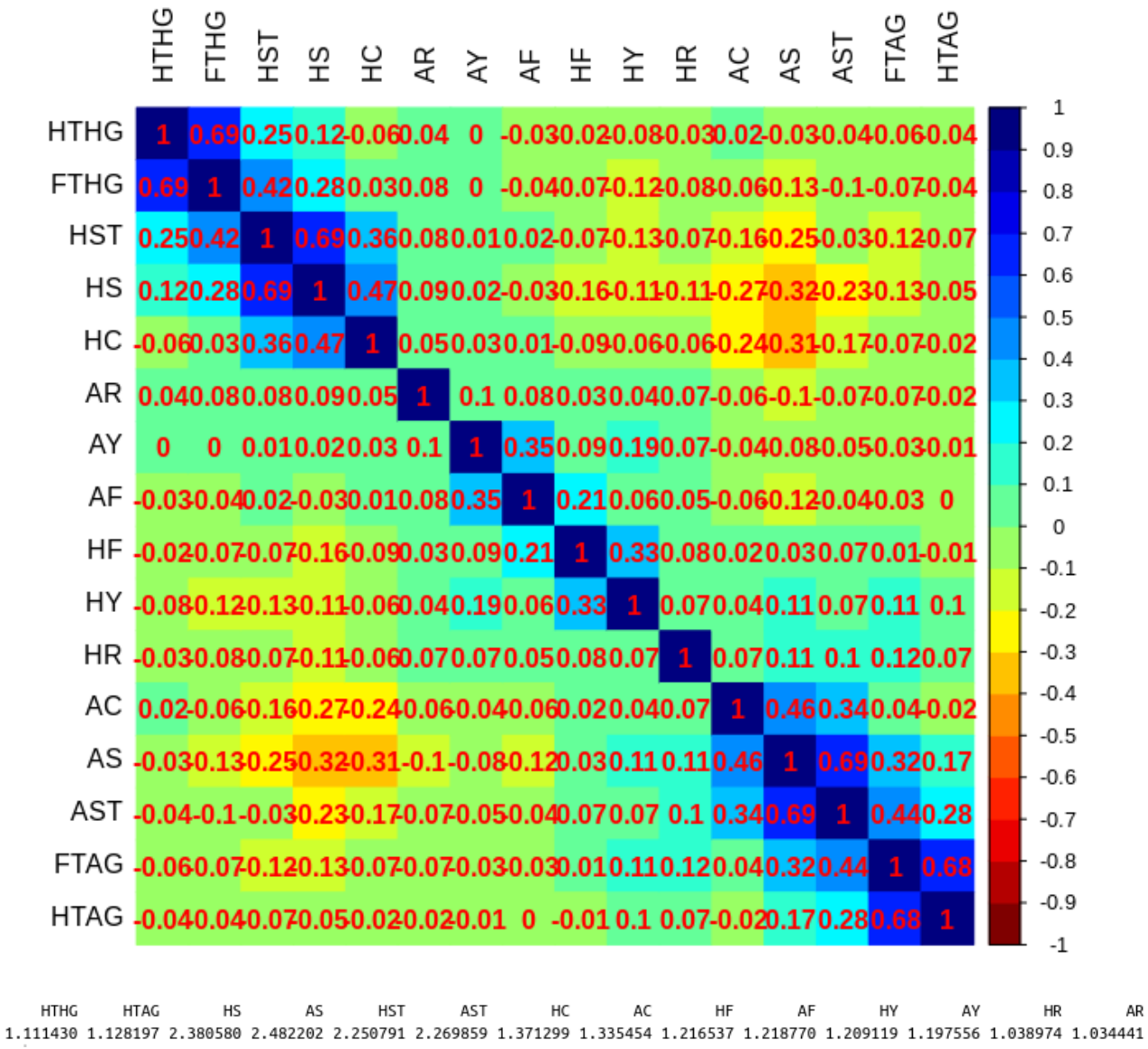
2. Проверить наличие мультиколлинеарности в собственных данных с помощью корреляционной матрицы и фактора инфляции дисперсии.

Код:

```
match_results <- read.csv(file.choose())
attach(match_results)
num_cols = colnames(match_results)
num_cols
num_cols <- num_cols[-c(1:4, 7, 10, 11)]  # non-numerical cols
numerical_results <- match_results[num_cols]  # get numerical cols
numerical_results <- na.omit(numerical_results)  # delete nans
M <- cor(numerical_results)
library(corrplot)
col4 <- colorRampPalette(c("#7F0000", "red", "#FF7F00", "yellow",
                           "#7FFF7F", "cyan", "#007FFF", "blue","#00007F"))
corrplot(M, method = "pie", col = col4(20), cl.length = 21,
         order = "AOE",  tl.col = "black", addCoef.col = "red")

corrplot(M, method = "color", col = col4(20), cl.length = 21,
         order = "AOE",  tl.col = "black", addCoef.col = "red")
vif(lm(FTHG - FTAG ~ HTHG + HTAG + HS + AS + HST + AST + HC + AC + HF + AF +
HY +
         AY + HR + AR,
       data = match_results))
```

Вывод:






```
HTHG     HTAG       HS       AS      HST      AST       HC       AC       HF       AF       HY       AY       HR       AR
1.111430 1.128197 2.380580 2.482202 2.250791 2.269859 1.371299 1.335454 1.216537 1.218770 1.209119 1.197556 1.038974 1.034441
```

Комментарий: попытка предсказать, кто выиграет матч не обвенчается успехом, так как этот ковариация между нашими атрибутами отсутствует, что доказывает и фактор инфляции дисперсии, большая часть параметров которого находятся между 1 и 2, что показывает отсутствие мультиколлинеарности.

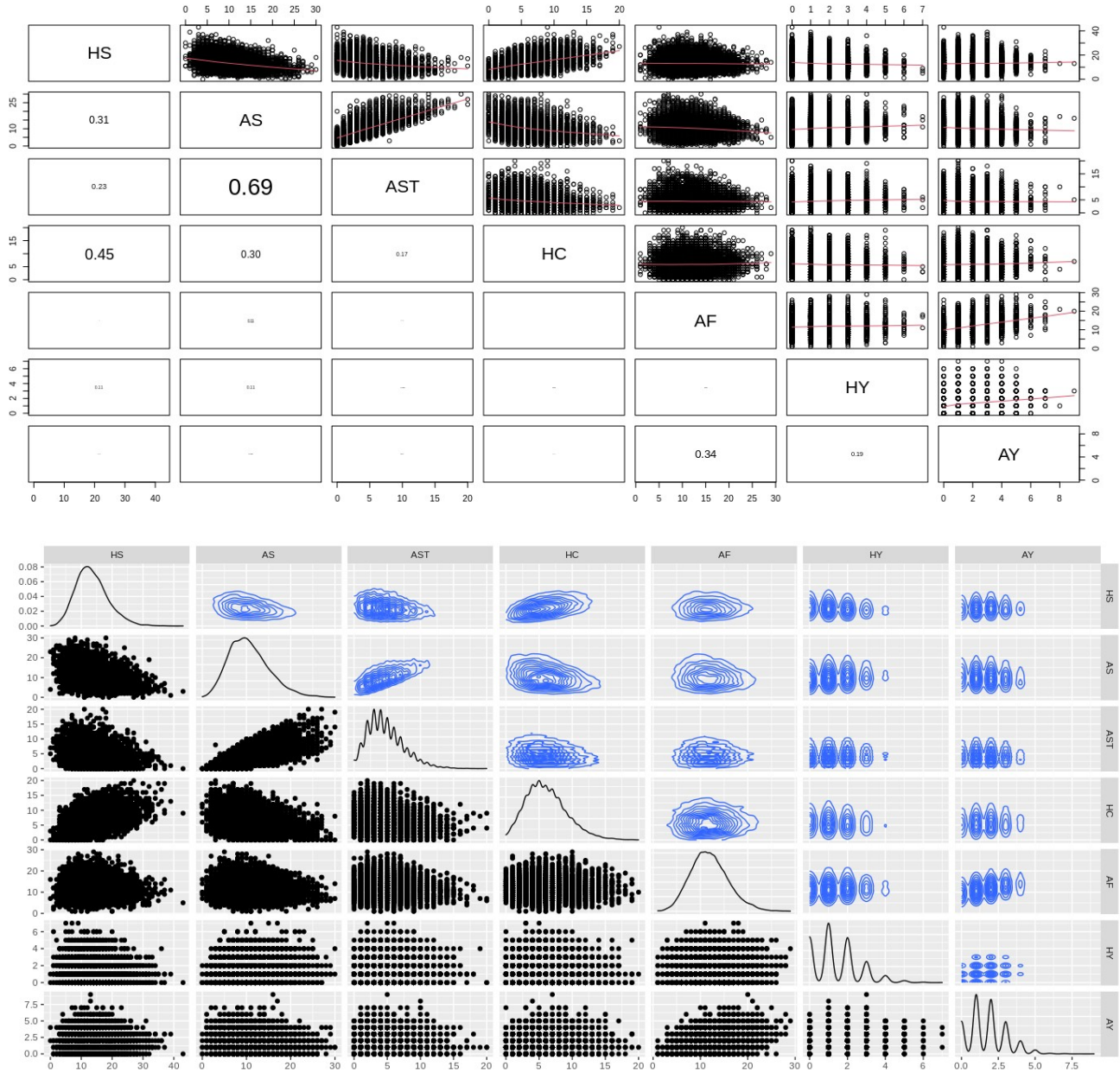
3. Изучить форму связи между переменными (количественными и категориальными) для выбранных данных.

Код:

```
numerical_results = numerical_results[num_cols[-c(8, 9)]]
# выкинем пару столбцов т.к. много слишком много столбцов
pairs(numerical_results, panel = panel.smooth) #со сглаживающей кривой

# Наносим на график коэффциенты корреляции Спирмена
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y, method = "spearman"))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep = "")
  # text(0.5, 0.5, txt)
```

```
  #Размер шрифта может зависеть от значения коэффициента корреляции
  if(missing(cex.cor))
    cex.cor <- 1.1/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(numerical_results, panel = panel.smooth, lower.panel = panel.cor)
library(lattice)
splom(cars)

library(GGally)
ggpairs(numerical_results,
        upper = list(continuous = "density", combo = "box"),
        lower = list(continuous = "points", combo = "dot"))
```

Вывод:

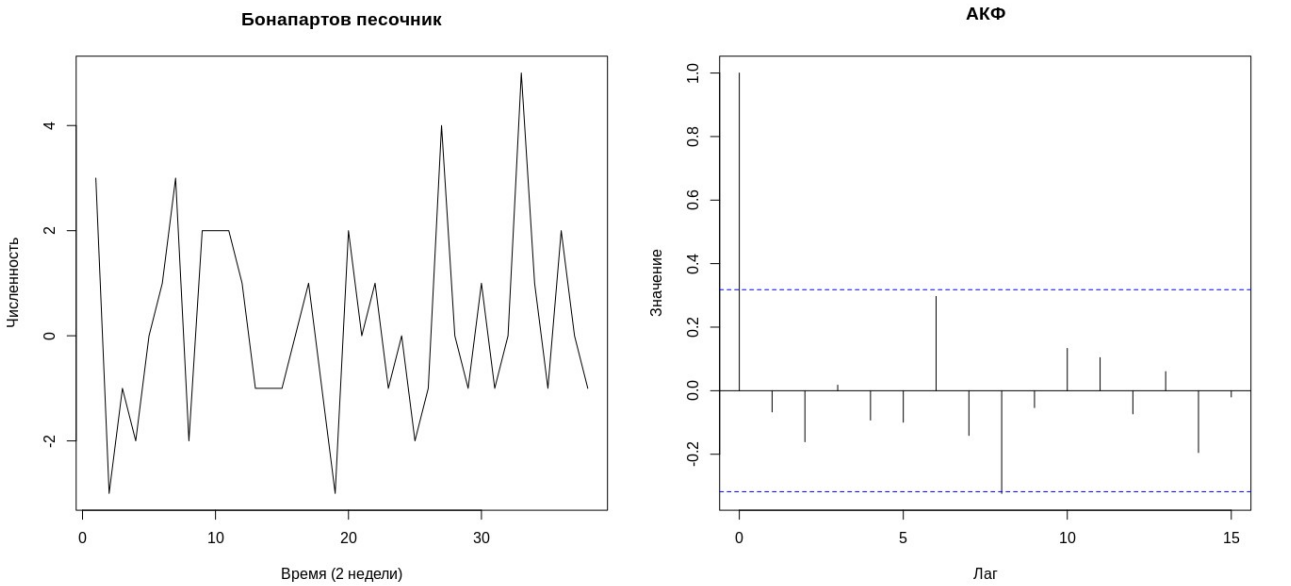


Комментарий: Еще одно доказательство того, что данные столбцы взаимно независимы.

Задание 4. Построить графики автокорреляционных функций для своих данных.
Код:

```
club = "Chelsea"
season = "2019-20"
Time = 38
club_matches = match_results[Season == season & (HomeTeam==club |
AwayTeam==club),]
club_matches
gds <- c()
for (row in 1:nrow(club_matches)) {
  x = match_results[row, ]
  if (x$HomeTeam == club) {
    gds <- c(gds, x$FTHG - x$FTAG)
  } else {
    gds <- c(gds, x$FTAG - x$FTHG)
  }
}
gds

layout(matrix(1:2, ncol = 2))
plot(1:Time, gds, type = "l", main="Бонапартов песочник",
     xlab = "Время (2 недели)", ylab = "Численность")
acf(gds, main = "АКФ", xlab = "Лаг", ylab = "Значение")
```

Вывод:



Комментарий: Построена АКФ для результатов матчей одной команды на протяжении определенного сезона. Как видно, не присутствует никакая цикличность, и результаты матчей независимы от предыдущих игр.

5. Исследовать зависимости в своих данных с помощью дисперсионного анализа.

Код:

```
M <- aov(FTHG - FTAG ~ HTHG + HTAG + HS + AS + HST + AST + HC + AC + HF + AF +
HY +
           AY + HR + AR)
summary(M)
```

```
              Df Sum Sq Mean Sq  F value   Pr(>F)
HTHG           1   7259    7259 5475.429  < 2e-16 ***
HTAG           1   4848    4848 3656.348  < 2e-16 ***
HS             1   1068    1068  805.530  < 2e-16 ***
AS             1    590     590  445.367  < 2e-16 ***
HST            1    361     361  272.304  < 2e-16 ***
AST            1    553     553  416.924  < 2e-16 ***
HC             1     54      54   40.424 2.16e-10 ***
AC             1     24      24   17.921 2.33e-05 ***
HF             1     17      17   12.898 0.000331 ***
AF             1      2       2    1.330 0.248802
HY             1     20      20   14.760 0.000123 ***
AY             1      3       3    2.609 0.106285
HR             1     62      62   46.743 8.69e-12 ***
AR             1     59      59   44.535 2.67e-11 ***
Residuals   7965  10560       1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2824 пропущенных наблюдений удалены
```

Комментарий: тут возникли небольшие проблемы с тем как воспринимать p-value. Правильно ли я понимаю что основная гипотеза состоит в том, что функция зависит от этого столбца, а альтернативная в том, что нет?