1. laboratorijska vježba

Multivarijatna analiza podataka

ak. god. 2021/2022

Verzija: 1.0

1. Uvod i upute za predaju

Cilj ove laboratorijske vježbe je primijeniti osnovne koncepte multivarijatne analize podataka, istražiti podatke te ispitati hipoteze. Preduvjet za rješavanje vježbe je osnovno znanje programskog jezika *R* i rad s *R Markdown* dokumentima. Sama vježba je koncipirana kao projekt u kojem istražujete i eksperimentirate koristeći dane podatke - ne postoji nužno samo jedan točan način rješavanja svakog podzadatka.

Rješavanje vježbe svodi se na čitanje uputa u tekstu ovog dokumenta, nadopunjavanje blokova kôda (možete dodavati i dodatne blokove kôda ukoliko je potrebno) i ispisivanje rezultata (u vidu ispisa iz funkcija, tablica i grafova). Vježbu radite samostalno, a svoje rješenje branite na terminima koji su vam dodijeljeni u kalendaru. Pritom morate razumjeti teorijske osnove u okviru onoga što je obrađeno na predavanjima i morate pokazati da razumijete sav kôd koji ste napisali.

Vaše rješenje potrebno je predati u sustav *Moodle* u obliku dvije datoteke:

1. Ovaj .Rmd dokument s Vašim rješenjem (naziva IME_PREZIME_JMBAG.rmd),

2. PDF ili HTML dokument kao izvještaj generiran iz vašeg .Rmd rješenja (također naziva IME_PREZIME_JMBAG).

Rok za predaju je **3. travnja 2022. u 23:59h**. Podsjećamo da bodovi iz laboratorijskih vježbi ulaze i u bodove na ispitnom roku, te da je za polaganje predmeta potrebno imati barem 50% ukupnih bodova iz laboratorijskih vježbi. **Nadoknade laboratorijskih vježbi neće biti organizirane.** Za sva dodatna pitanja svakako se javite na email adresu predmeta: *map@fer.hr*.

2. Podatkovni skup

Podatkovni skup koji će biti razmatran u vježbi sadrži bodove studenata na jednom fakultetskom kolegiju. Svakom studentu upisani su bodovi iz dviju laboratorijskih vježbi (**LAB**), pet zadataka međuispita (**MI**), pet zadataka završnog ispita (**ZI**), pet zadataka ispitnog roka (**IR**) i kojoj grupi predavanja pripadaju (**Grupa**).

Studenti mogu položiti kolegij kontinuiranim putem ili na ispitnom roku. Kontinuirani put sastoji se od bodova s laboratorijskih vježbi, međuispita i završnog ispita. Kronološki, 1. laboratorijska vježba održana je prije međuispita, dok je 2. laboratorijska vježba održana između međuispita i završnog ispita. Ispitni rok održan je nakon završnog ispita. Ako student polaže predmet na ispitnom roku, gledaju se samo bodovi s ispitnog roka. Ukupan broj bodova je 100, a bodovi su raspodijeljeni na sljedeći način:

- Kontinuirana nastava:
  - **LAB**: 20 bodova (0-10 svaka vježba)
  - **MI** : 40 bodova (0-8 svaki zadatak)
  - **ZI** : 40 bodova (0-8 svaki zadatak)
- Ispitni rok:
  - **IR** : 100 bodova (0-20 svaki zadatak)

Za prolazak kolegija potrebno je skupiti **više** od 50 bodova i izaći na obje laboratorijske vježbe (izlazak na vježbe nužan je uvjet i za polaganje ispitnog roka, iako se bodovi ne prenose). Ako student nije pristupio pripadajućem ispitu/laboratorijskoj vježbi, nije upisan podatak (što nije isto kao i 0 bodova).

3. Priprema podataka i eksploratorna analiza

U ovom dijelu vježbe potrebno je učitati podatke i napraviti osnovnu eksploratornu analizu podataka.

3.1 Učitavanje podataka

Učitajte podatkovni skup iz datoteke *studenti.csv* i pripremite podatke za analizu. Pritom obratite pozornost na sljedeće:

- Provjerite jesu li sve varijable očekivanog tipa,
- Provjerite jesu li vrijednosti unutar zadanog raspona (s obzirom na gore opisano bodovanje),
- Provjerite zadovoljavaju li bodovi gore opisane uvjete predmeta,
- Za nedostajuće podatke ispitajte jesu li opravdani te odaberite i primijenite tehniku upravljanja nedostajućim podatcima.

Nakon što su podatci pripremljeni, analizirajte i ispišite deksriptivne statistike varijabli.

1) Provjera tipova varijabli:

```
df <- read.csv('./studenti[1].csv')
head(df)
```

```
##   MI_1 MI_2 MI_3 MI_4 MI_5 LAB_1 ZI_1 ZI_2 ZI_3 ZI_4 ZI_5 LAB_2 IR_1 IR_2 IR_3
## 1  7.5  6.5  4.0  3.0    2     8  4.5  6.5  6.0  4.0    3     2   NA   NA   NA
## 2  7.5  3.5  4.0  4.0    0     7  8.0  6.0  4.0  2.0  3.5     5   NA   NA   NA
## 3  6.0  4.5  4.5  4.5  0.5   5.5  6.5  6.5  3.5  2.5    2     2 15.0   18 14.5
## 4  5.5  6.5  2.5  3.0    0   4.5  3.5  6.5  2.5  2.5  1.5     1 19.0   16 14.0
## 5  6.0  2.0  3.5  3.5  3.5   7.5  3.5  5.0  4.0  2.5  2.5     3 18.5   20 12.5
## 6  8.0  5.0  3.5  2.5  4.5   8.5  6.0  6.0  3.0  2.0  2.5   5.5   NA   NA   NA
##   IR_4 IR_5 Grupa
## 1   NA   NA     2
## 2   NA   NA     1
## 3 16.0  7.0     2
## 4  7.5  7.0     2
## 5 14.0  7.5     3
## 6   NA   NA     2
```

1) Izmjene krivo unesenih vrijednosti:

```
# MI_5, ZI_5, LAB_1, LAB_2 su problematicni jer se radi o characterima
print(sapply(df, typeof))
```

```
##        MI_1        MI_2        MI_3        MI_4        MI_5       LAB_1
##    "double"    "double"    "double"    "double" "character" "character"
##        ZI_1        ZI_2        ZI_3        ZI_4        ZI_5       LAB_2
##    "double"    "double"    "double"    "double" "character" "character"
##        IR_1        IR_2        IR_3        IR_4        IR_5       Grupa
##    "double"    "double"    "double"    "double"    "double"   "integer"
```

```
print(unique(df$LAB_1))
```

```
##  [1] "8"    "7"    "5.5"  "4.5"  "7.5"  "8.5"  "6"    "5"    "6.5"  "4"
## [11] "9"    "9.5"  "NULL"
```

```r
print(unique(df$LAB_2))
```

```
##  [1] "2"    "5"    "1"    "3"    "5.5"  "3.5"  "2.5"  "4"    "1.5"  "4.5"
## [11] "0.5"  "NULL" "6"    NA
```

```r
print(unique(df$MI_5))
```

```
##  [1] "2"    "0"    "0.5"  "3.5"  "4.5"  "5.5"  "3"    "7"    "5"    "6"
## [11] "4"    "1.5"  "6.5"  "1"    "2.5"  "0.0/" "7.5"  "8"
```

```r
print(unique(df$ZI_5))
```

```
##  [1] "3"    "3.5"  "2"    "1.5"  "2.5"  "0.5"  "4.5"  "0"    "1"    "4"
## [11] "5.5"  "0.5p"
```

```r
df$LAB_1[df$LAB_1 == "NULL"] <- NA

df$LAB_2[df$LAB_2 == "NULL"] <- NA

df$MI_5[df$MI_5 == "0.0/"] <- 0.0

df$ZI_5[df$ZI_5 == "0.5p"] <- 0.5

print(unique(df$LAB_1))
```

```
##  [1] "8"   "7"   "5.5" "4.5" "7.5" "8.5" "6"   "5"   "6.5" "4"   "9"   "9.5"
## [13] NA
```

```r
print(unique(df$LAB_2))
```

```
##  [1] "2"   "5"   "1"   "3"   "5.5" "3.5" "2.5" "4"   "1.5" "4.5" "0.5" NA
## [13] "6"
```

```r
print(unique(df$MI_5))
```

```
##  [1] "2"   "0"   "0.5" "3.5" "4.5" "5.5" "3"   "7"   "5"   "6"   "4"   "1.5"
## [13] "6.5" "1"   "2.5" "7.5" "8"
```

```r
print(unique(df$ZI_5))
```

```
##  [1] "3"   "3.5" "2"   "1.5" "2.5" "0.5" "4.5" "0"   "1"   "4"   "5.5"
```

Pretvorba tipova:

```r
df$LAB_1 <- as.numeric(df$LAB_1)
df$LAB_2 <- as.numeric(df$LAB_2)
df$MI_5 <- as.numeric(df$MI_5)
df$ZI_5 <- as.numeric(df$ZI_5)

print(sapply(df, typeof))
```

```
##      MI_1      MI_2      MI_3      MI_4      MI_5     LAB_1      ZI_1      ZI_2
##  "double"  "double"  "double"  "double"  "double"  "double"  "double"  "double"
##      ZI_3      ZI_4      ZI_5     LAB_2      IR_1      IR_2      IR_3      IR_4
##  "double"  "double"  "double"  "double"  "double"  "double"  "double"  "double"
##      IR_5     Grupa
##  "double" "integer"
```

2) Provjera raspona vrijednosti:

```r
df_range = t(apply(df, 2, range, na.rm=TRUE))
colnames(df_range) <- c('MIN', 'MAX')
df_range
```

```
##        MIN  MAX
## MI_1   4.0  8.0
## MI_2   0.0 18.0
## MI_3   0.0  8.0
## MI_4   0.5  7.0
## MI_5   0.0  8.0
## LAB_1  4.0  9.5
## ZI_1  -3.0  8.0
## ZI_2   3.0  8.0
## ZI_3   0.0  8.0
## ZI_4   0.0  5.5
## ZI_5   0.0  5.5
## LAB_2  0.5  6.0
## IR_1   0.0 20.0
## IR_2   0.0 20.0
## IR_3   0.0 18.5
## IR_4   0.0 20.0
## IR_5   0.0 11.5
## Grupa  1.0  3.0
```

```r
print(df$MI_2[df$MI_2 > 10])
```

```
## [1] 18
```

```r
print(df$ZI_1[df$ZI_1 < 0])
```

```
## [1] -3
```

```r
df<-df[(df$MI_2 <= 10 & df$ZI_1 > 0),]

# posto se radi o samo dva primjera nije toliko bitno, ali moguce i da su pogreske bile:
# 18 -> 8
# -3 -> 3
```

```r
df_range = t(apply(df, 2, range, na.rm=TRUE))
colnames(df_range) <- c('MIN', 'MAX')
df_range
```

```
##       MIN  MAX
## MI_1  4.0  8.0
## MI_2  0.0  8.0
## MI_3  0.0  8.0
## MI_4  0.5  7.0
## MI_5  0.0  8.0
## LAB_1 4.0  9.5
## ZI_1  0.5  8.0
## ZI_2  3.0  8.0
## ZI_3  0.0  8.0
## ZI_4  0.0  5.5
## ZI_5  0.0  5.5
## LAB_2 0.5  6.0
## IR_1  0.0 20.0
```

```
## IR_2  0.0 20.0
## IR_3  0.0 18.5
## IR_4  0.0 20.0
## IR_5  0.0 11.5
## Grupa 1.0  3.0
```

Kod MI_2 varijable je postoji problematična vrijednost 18, koja bi trebala biti manja ili jednaka 10. Druga problematična vrijednost je vrijednost varijable ZI_1 koja je negativna, a trebala bi biti veća ili jednaka 0.

3) Provjera zadovoljavaju li bodovi kriterije:

```
# ne moze se dogoditi da netko ima bodove iz ispita ako nema oba labosa
print(length(df[is.na(df$LAB_1) & is.na(df$LAB_2),]))
```

```
## [1] 18
```

```
df = df[!is.na(df$LAB_1) & !is.na(df$LAB_2),]
```

Više smisla ima gledati deskriptivne statistike tako da se ignoriraju NA vrijednosti pa ću to napraviti prije zamjene nedostajućih podataka.

```
summary(df)
```

```
##       MI_1            MI_2            MI_3           MI_4            MI_5
##  Min.   :4.000   Min.   :0.000   Min.   :0.00   Min.   :0.500   Min.   :0.00
##  1st Qu.:6.500   1st Qu.:4.500   1st Qu.:3.50   1st Qu.:3.500   1st Qu.:1.50
##  Median :7.000   Median :6.000   Median :5.00   Median :4.000   Median :3.00
##  Mean   :6.918   Mean   :5.828   Mean   :4.94   Mean   :4.011   Mean   :3.04
##  3rd Qu.:7.500   3rd Qu.:7.500   3rd Qu.:6.50   3rd Qu.:4.500   3rd Qu.:4.50
##  Max.   :8.000   Max.   :8.000   Max.   :8.00   Max.   :7.000   Max.   :8.00
##
##      LAB_1            ZI_1           ZI_2            ZI_3            ZI_4
##  Min.   :4.000   Min.   :0.50   Min.   :3.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:6.500   1st Qu.:4.50   1st Qu.:5.500   1st Qu.:2.500   1st Qu.:2.500
##  Median :7.000   Median :6.00   Median :6.000   Median :4.000   Median :3.000
##  Mean   :6.991   Mean   :5.84   Mean   :5.991   Mean   :4.014   Mean   :3.003
##  3rd Qu.:7.500   3rd Qu.:7.50   3rd Qu.:6.500   3rd Qu.:5.500   3rd Qu.:3.500
##  Max.   :9.500   Max.   :8.00   Max.   :8.000   Max.   :8.000   Max.   :5.500
##
##       ZI_5           LAB_2            IR_1            IR_2           IR_3
##  Min.   :0.000   Min.   :0.500   Min.   : 0.00   Min.   : 0.0   Min.   : 0.00
##  1st Qu.:1.500   1st Qu.:2.500   1st Qu.:13.50   1st Qu.:12.5   1st Qu.:13.50
##  Median :2.000   Median :3.000   Median :15.25   Median :14.5   Median :14.50
##  Mean   :2.008   Mean   :3.003   Mean   :15.41   Mean   :14.1   Mean   :14.43
##  3rd Qu.:2.500   3rd Qu.:3.500   3rd Qu.:18.00   3rd Qu.:16.0   3rd Qu.:15.62
##  Max.   :5.500   Max.   :6.000   Max.   :20.00   Max.   :20.0   Max.   :18.50
##                                  NA's   :399     NA's   :399    NA's   :399
##       IR_4            IR_5           Grupa
##  Min.   : 0.000   Min.   : 0.00   Min.   :1.000
##  1st Qu.: 7.875   1st Qu.: 5.00   1st Qu.:1.000
##  Median :11.000   Median : 6.50   Median :2.000
##  Mean   :11.073   Mean   : 6.26   Mean   :2.004
##  3rd Qu.:14.125   3rd Qu.: 7.50   3rd Qu.:3.000
##  Max.   :20.000   Max.   :11.50   Max.   :3.000
##  NA's   :399      NA's   :399
```

4) Zamjena nedostajućih podataka:

```
df['take_exam'] = ifelse(
  !is.na(df$IR_1) & !is.na(df$IR_2) & !is.na(df$IR_3) & !is.na(df$IR_4) & !is.na(df$IR_5),
  1,
  0
)

head(df['take_exam'])
```

```
##   take_exam
## 1         0
## 2         0
## 3         1
## 4         1
## 5         1
## 6         0
```

Za NA vrijednosti na IR, MI i ZI mi najviše smisla ima zamijeniti nulama jer je na ovom predmetu identično ne izaći na ispit i predati potpuno prazan ispit. Međutim, ovo će definitivno utjecati na deskriptivnu statistiku.

```
df[is.na(df)] <- 0
```

3.2 Korelacijska analiza

Razmotrimo studente koji su predmet položili kontinuirano. Izračunajte i vizualizirajte matricu korelacije za njihove bodove na nastavnim aktivnostima. Ponovite isto za studente koji su izašli na ispitni rok. Razmislite o zavisnosti različitih nastavnih aktivnosti koje vidite iz ovih korelacijskih matrica.

```
#  Vaš kôd ovdje
df_tmp = df[df$take_exam == 0,]

# suma bodova > 50
df_passed_cont = df_tmp[((df_tmp$LAB_1 + df_tmp$LAB_2) +
  (df_tmp$MI_1 + df_tmp$MI_2 + df_tmp$MI_3 + df_tmp$MI_4 + df_tmp$MI_5) +
  (df_tmp$ZI_1 + df_tmp$ZI_2 + df_tmp$ZI_3 + df_tmp$ZI_4 + df_tmp$ZI_5)) >= 50,]

head(df_passed_cont)
```

```
##   MI_1 MI_2 MI_3 MI_4 MI_5 LAB_1 ZI_1 ZI_2 ZI_3 ZI_4 ZI_5 LAB_2 IR_1 IR_2 IR_3
## 1  7.5  6.5  4.0  3.0  2.0   8.0  4.5  6.5  6.0    4  3.0   2.0    0    0    0
## 2  7.5  3.5  4.0  4.0  0.0   7.0  8.0  6.0  4.0    2  3.5   5.0    0    0    0
## 6  8.0  5.0  3.5  2.5  4.5   8.5  6.0  6.0  3.0    2  2.5   5.5    0    0    0
## 7  7.0  4.0  5.0  4.5  3.5   8.0  5.5  7.0  6.0    2  0.5   3.0    0    0    0
## 8  6.0  7.5  7.5  3.5  5.5   6.0  8.0  7.5  6.0    2  3.0   3.0    0    0    0
## 9  7.5  6.0  4.0  4.5  3.5   7.5  7.0  5.0  1.5    4  1.5   3.5    0    0    0
##   IR_4 IR_5 Grupa take_exam
## 1    0    0     2         0
## 2    0    0     1         0
## 6    0    0     2         0
## 7    0    0     2         0
## 8    0    0     2         0
## 9    0    0     2         0
```

```
library(ggplot2)
```

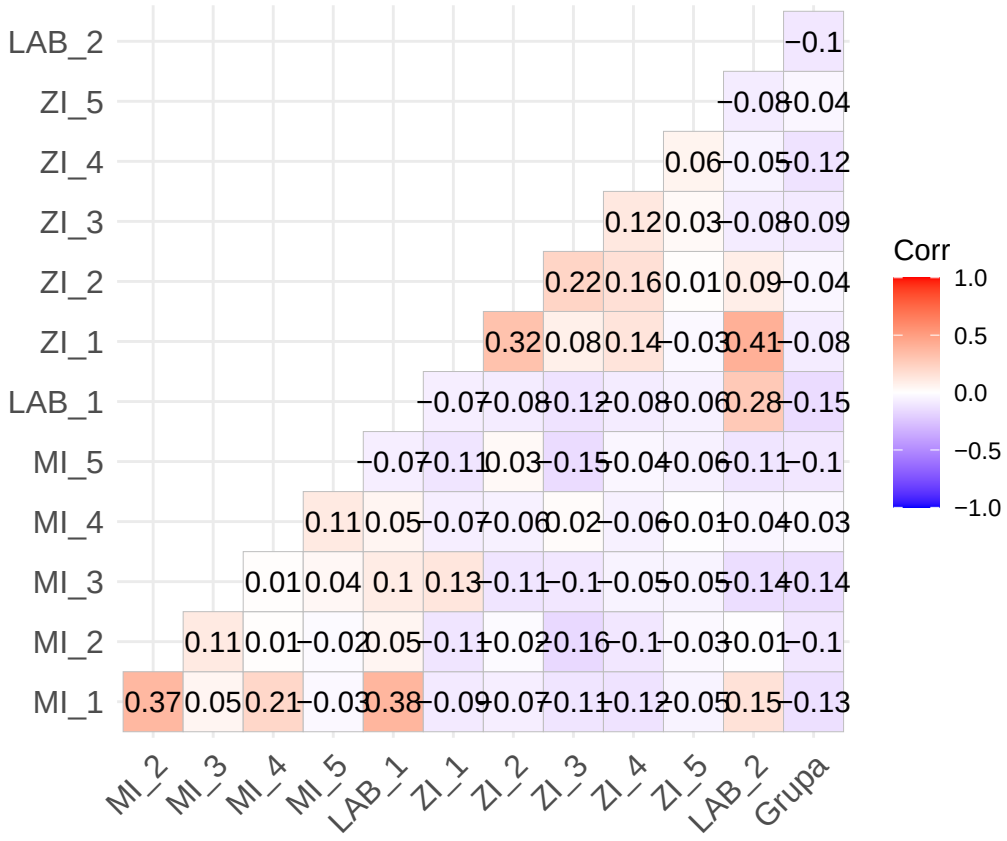
```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.0.5
```

```
ggcorrplot(
  cor(df_passed_cont, method='pearson'),
  type='lower',
  digits=2,
  lab=TRUE,
)
```

```
## Warning in cor(df_passed_cont, method = "pearson"): the standard deviation is
## zero
```



Dio koji se odnosi na studente na ispitnom roku:

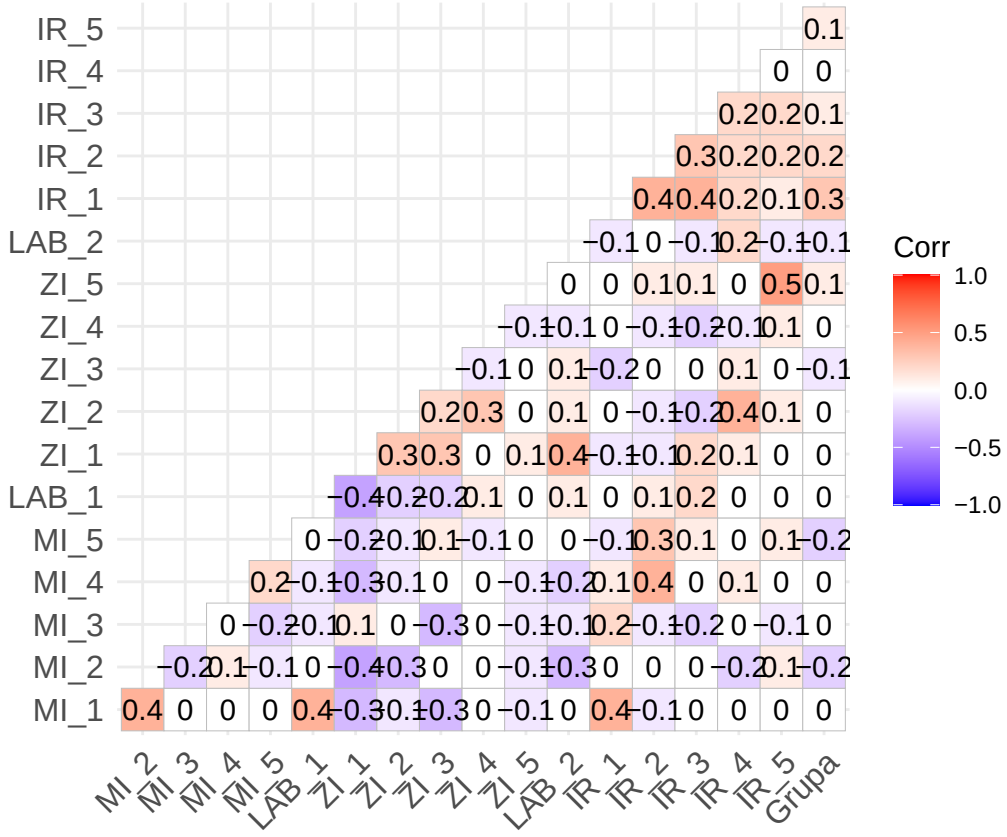
```
#  Vaš kôd ovdje
df_exam = df[df$take_exam == 1,]

head(df_exam)
```

```
##    MI_1 MI_2 MI_3 MI_4 MI_5 LAB_1 ZI_1 ZI_2 ZI_3 ZI_4 ZI_5 LAB_2 IR_1 IR_2 IR_3
## 3   6.0  4.5  4.5  4.5  0.5   5.5  6.5  6.5  3.5  2.5  2.0   2.0 15.0 18.0 14.5
## 4   5.5  6.5  2.5  3.0  0.0   4.5  3.5  6.5  2.5  2.5  1.5   1.0 19.0 16.0 14.0
## 5   6.0  2.0  3.5  3.5  3.5   7.5  3.5  5.0  4.0  2.5  2.5   3.0 18.5 20.0 12.5
## 10  6.0  5.0  1.0  3.5  3.0   5.0  5.0  4.5  3.5  3.0  4.5   3.5 14.0 15.5 14.0
## 16  5.5  3.5  4.0  3.5  4.0   7.5  4.0  6.0  1.5  2.5  3.5   3.0 11.5 16.0 14.0
## 26  6.5  5.0  3.5  5.5  1.5   8.0  0.5  5.5  0.5  3.5  1.0   2.0 17.0 14.0 16.5
##    IR_4 IR_5 Grupa take_exam
## 3  16.0  7.0     2         1
```

```
## 4   7.5  7.0     2         1
## 5  14.0  7.5     3         1
## 10  8.0 10.5     2         1
## 16 12.5  9.5     2         1
## 26 15.0  3.5     2         1
```

```
ggcorrplot(
  cor(df_exam, method='pearson'),
  type='lower',
  digits=1,
  lab=TRUE,
)
```

```
## Warning in cor(df_exam, method = "pearson"): the standard deviation is zero
```



Prikažite upareni graf za zadatke s ispitnog roka. Na dijagonalama prikažite empirijsku distribuciju podataka, a na elementima izvan dijagonala prikažite grafove raspršenja za parove varijabli. Razmislite o karakteristikama grafova i razmislite postoje li primjeri koji odskaču od ostalih.

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.0.5
```

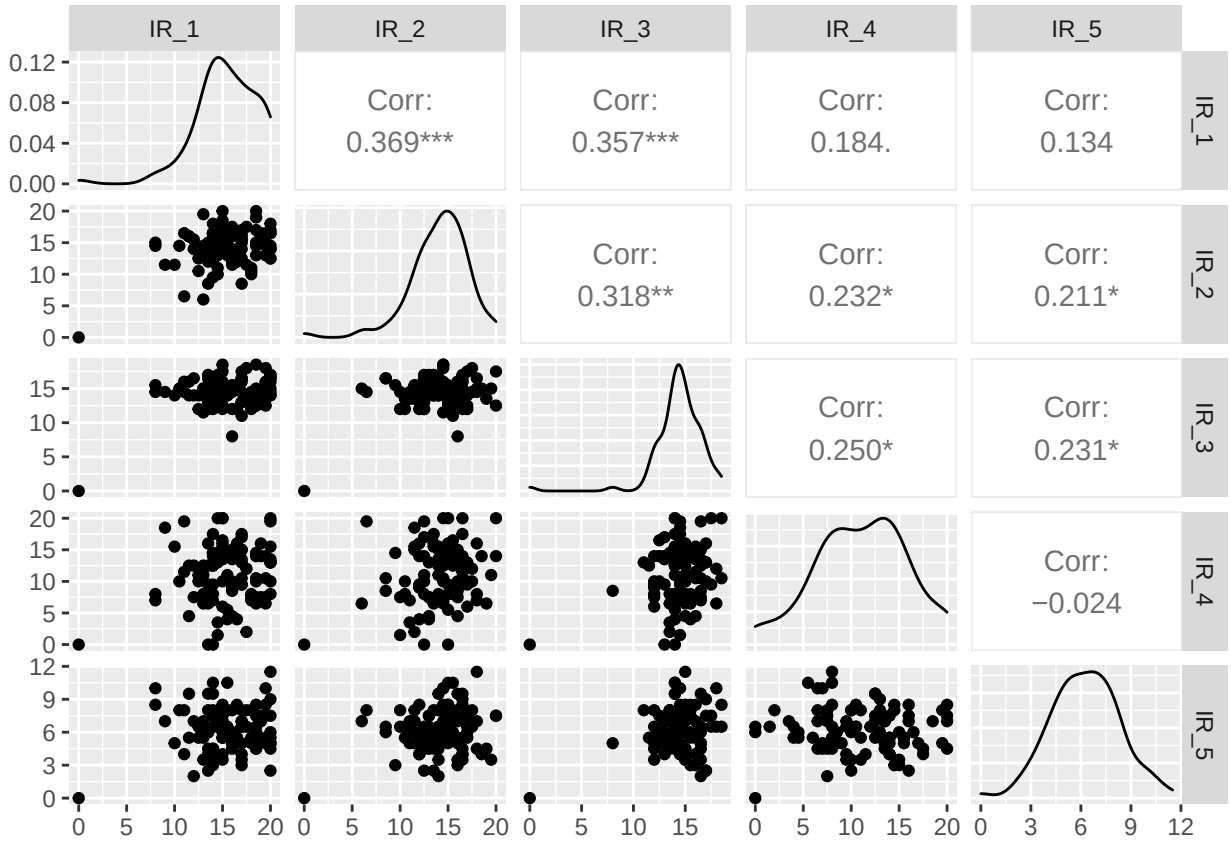
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
ggpairs(df_exam, columns=c('IR_1', 'IR_2', 'IR_3', 'IR_4', 'IR_5'), progress=F)
```



3.3 Statistička udaljenost

Izračunajte procjene vektora očekivanja i matrice kovarijance za zadatke s ispitnog roka, kao i statističke udaljenosti svih primjera u odnosu na procijenjeno očekivanje i kovarijancu. Ispitajte postoje li stršeće vrijednosti koje su statistički značajne.

```
df_exam_problems = df_exam[c('IR_1', 'IR_2','IR_3','IR_4','IR_5')]

head(df_exam_problems)
```

```
##    IR_1 IR_2 IR_3 IR_4 IR_5
## 3  15.0 18.0 14.5 16.0  7.0
## 4  19.0 16.0 14.0  7.5  7.0
```

```
## 5  18.5 20.0 12.5 14.0  7.5
## 10 14.0 15.5 14.0  8.0 10.5
## 16 11.5 16.0 14.0 12.5  9.5
## 26 17.0 14.0 16.5 15.0  3.5
```

```
mean_df = colMeans(df_exam_problems)
cov_df = cov(df_exam_problems)

print(mean_df)
```

```
##      IR_1      IR_2      IR_3      IR_4      IR_5
## 15.406250 14.104167 14.432292 11.072917  6.260417
```

```
print(cov_df)
```

```
##            IR_1     IR_2     IR_3       IR_4       IR_5
## IR_1 10.8911184 3.738816 2.727796  2.8805921  0.9141447
## IR_2  3.7388158 9.441667 2.262390  3.3765351  1.3331140
## IR_3  2.7277961 2.262390 5.350630  2.7418311  1.0993969
## IR_4  2.8805921 3.376535 2.741831 22.5156798 -0.2323465
## IR_5  0.9141447 1.333114 1.099397 -0.2323465  4.2419956
```

```
stat_dist = mahalanobis(df_exam_problems, mean_df, cov_df)
distr <- pchisq(stat_dist, df=ncol(df_exam_problems), lower.tail=FALSE)

print(length(stat_dist))
```

```
## [1] 96
```

```
print(length(stat_dist[distr > 0.01]))
```

```
## [1] 94
```

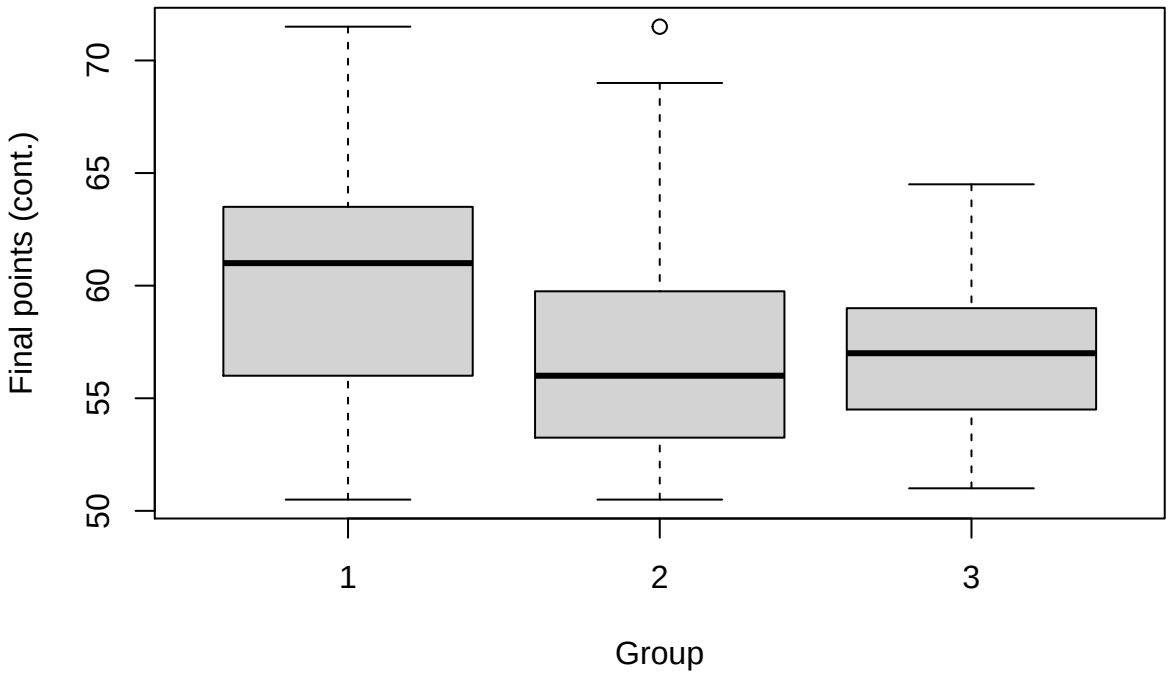
4. Analiza podataka

4.1 Vizualizacija i deskriptivna statistika

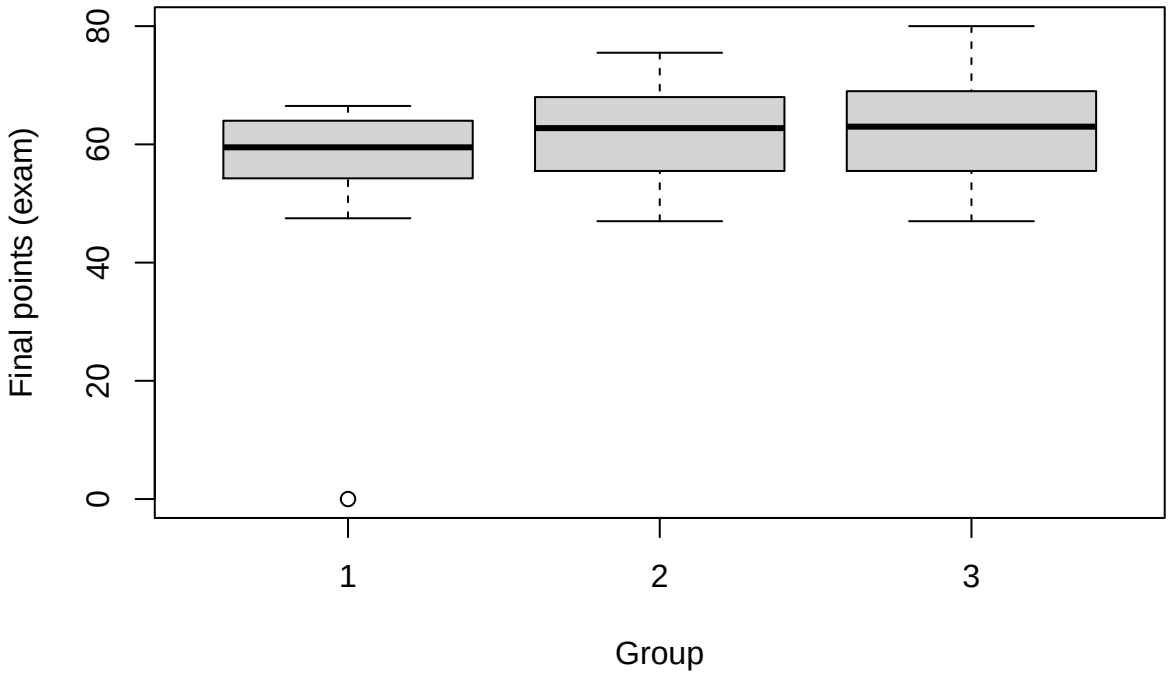
Analizirajte u podatcima sljedeća istraživačka pitanja, koristeći odgovarajuće vizualizacije i deskriptivne statistike ili druge tehnike (dodatno možete provesti i statistički test - nije obavezno).

- Imaju li grupe utjecaj na ukupne bodove iz kontinuirane nastave (postoje li grupe koje su uspješnije od ostalih)? Vrijedi li isto za bodove na roku?

```
# Vaš kôd ovdje

# boxplot s grupama kont.

sum_pts = df_passed_cont$MI_1 + df_passed_cont$MI_2 + df_passed_cont$MI_3 + df_passed_cont$MI_4 +
  df_passed_cont$MI_5 + df_passed_cont$ZI_1 + df_passed_cont$ZI_2 + df_passed_cont$ZI_3 +
  df_passed_cont$ZI_4 + df_passed_cont$ZI_5 + df_passed_cont$LAB_1 + df_passed_cont$LAB_2

boxplot(sum_pts ~ df_passed_cont$Grupa,
  xlab ="Group",
  ylab = "Final points (cont.)",
)
```

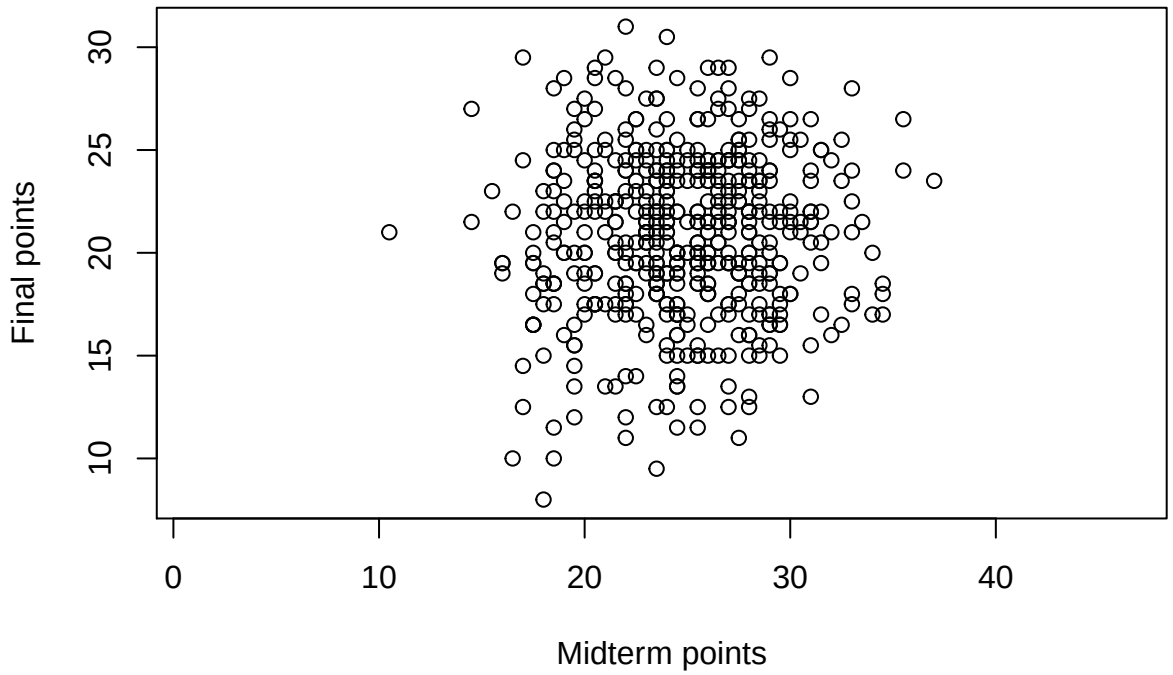



```
# boxplot s grupama rok

sum_pts2 = df_exam$IR_1 + df_exam$IR_2 + df_exam$IR_3 + df_exam$IR_4 + df_exam$IR_5

boxplot(sum_pts2 ~ df_exam$Grupa,
  xlab ="Group",
  ylab = "Final points (exam)",
)
```



- Postoji li povezanost između uspjeha studenata na međuispitu i završnom ispitu (vrijedi li da su uspješniji studenti na MI ujedno uspješniji i na ZI)?

```r
# Vaš kôd ovdje

# sumirati sve mi zadatke, sve zi zadatke i scatter plottat
df$mi <- df$MI_1 + df$MI_2 + df$MI_3 + df$MI_4 + df$MI_5
df$zi <- df$ZI_1 + df$ZI_2 + df$ZI_3 + df$ZI_4 + df$ZI_5

plot(df$mi,
  df$zi,
  asp=1,
  xlab='Midterm points',
  ylab='Final points'
)
```

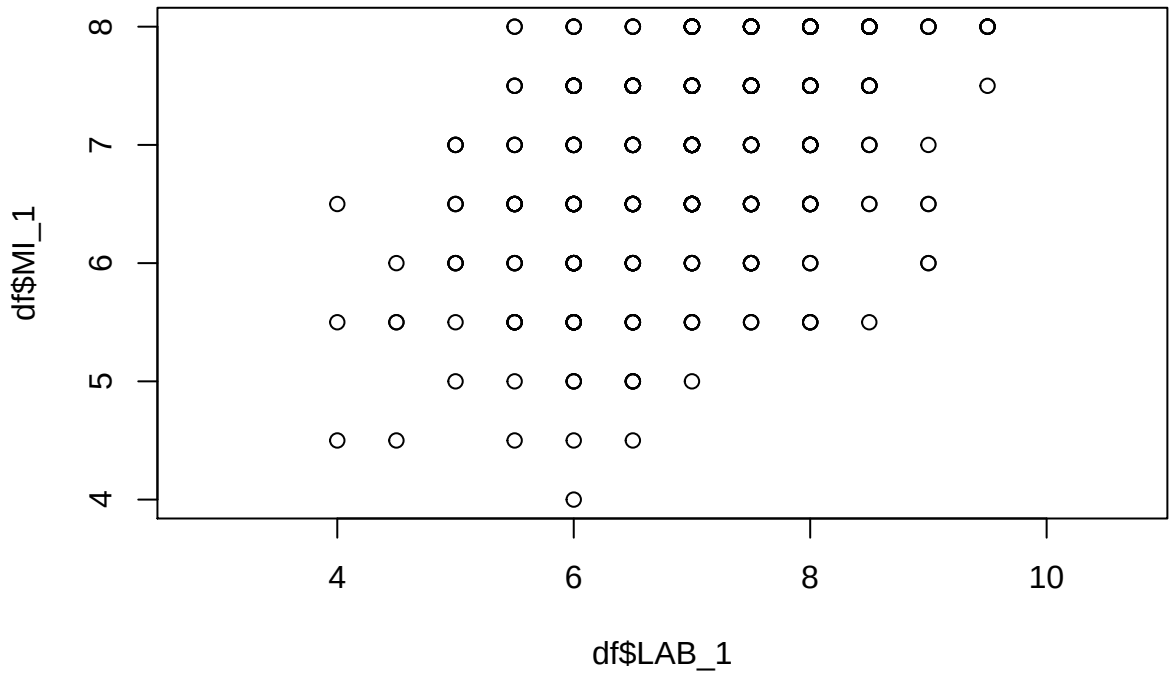



```r
print(cor(df$mi, df$zi))
```

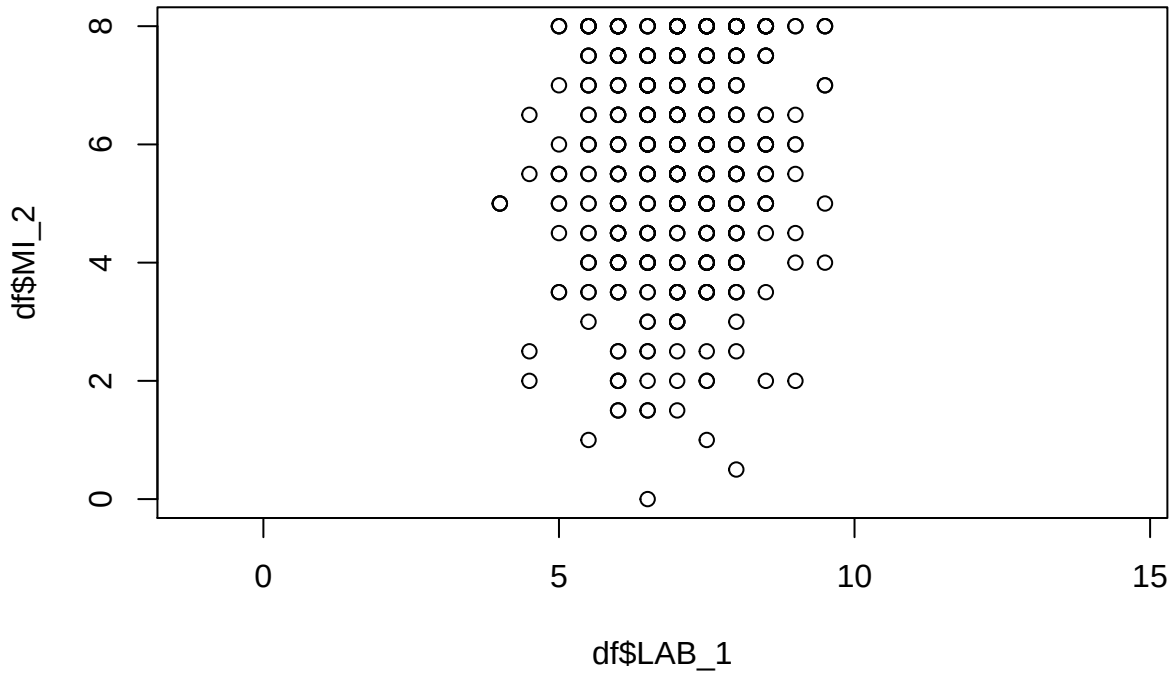
```
## [1] 0.06203261
```

- Postoji li povezanost između uspjeha studenata na nekim zadatcima na ispitima i pojedinim laboratorijskim vježbama? Razmislite koji su mogući uzroci ovakvih zavisnosti, ako postoje.

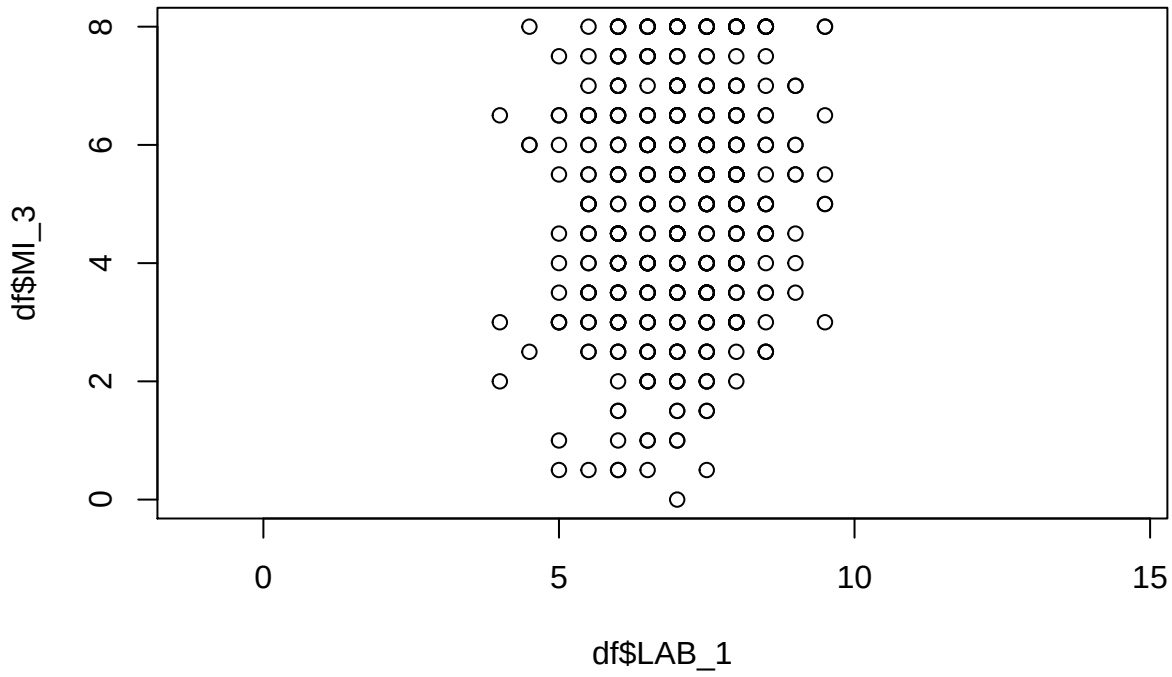
```r
# Vaš kôd ovdje

# scatterplot labosa i zadataka, eventualno korelacije

plot(df$LAB_1, df$MI_1, asp=1)
```



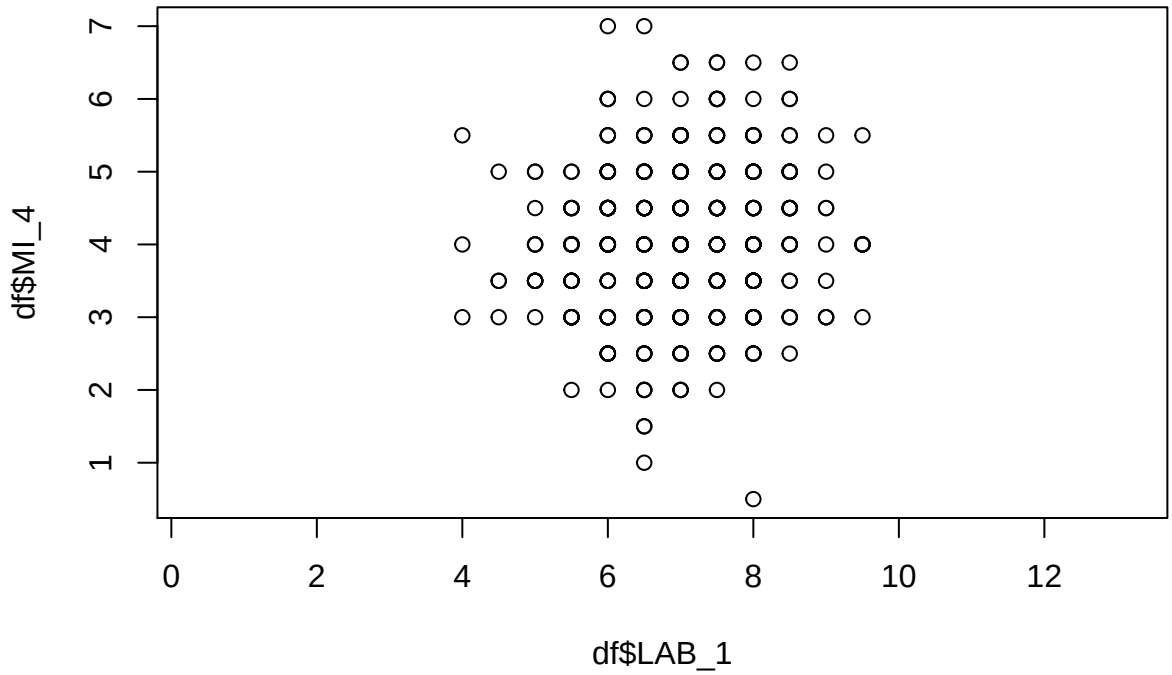

```
plot(df$LAB_1, df$MI_2, asp=1)
```



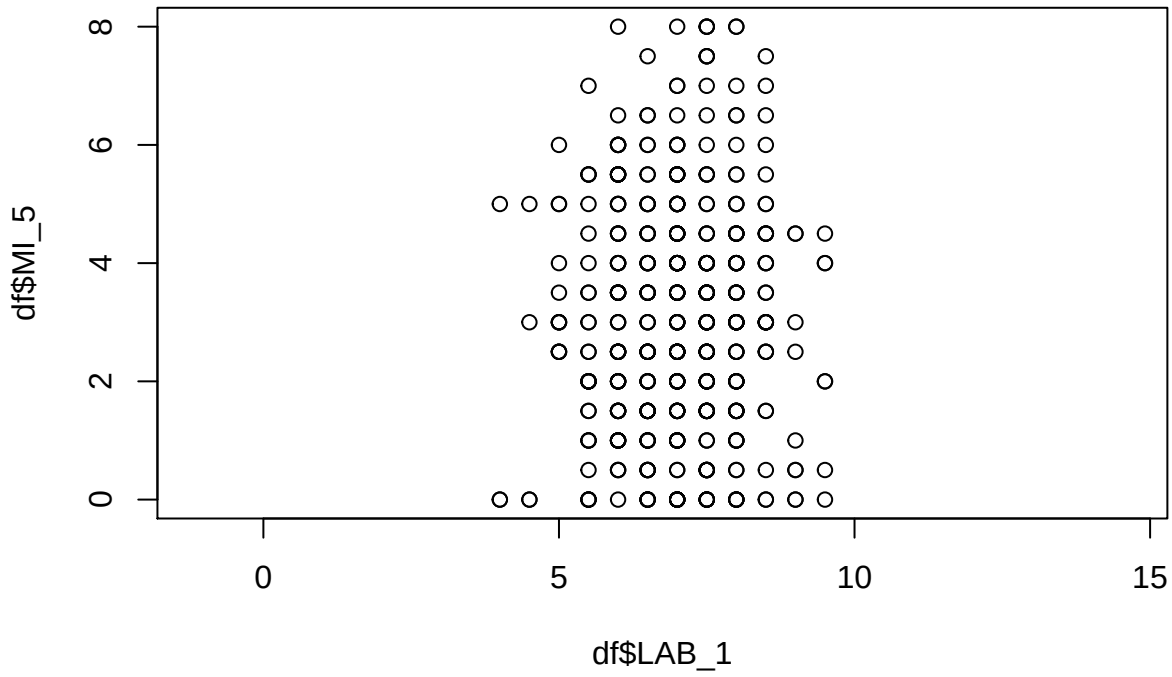

```
plot(df$LAB_1, df$MI_3, asp=1)
```



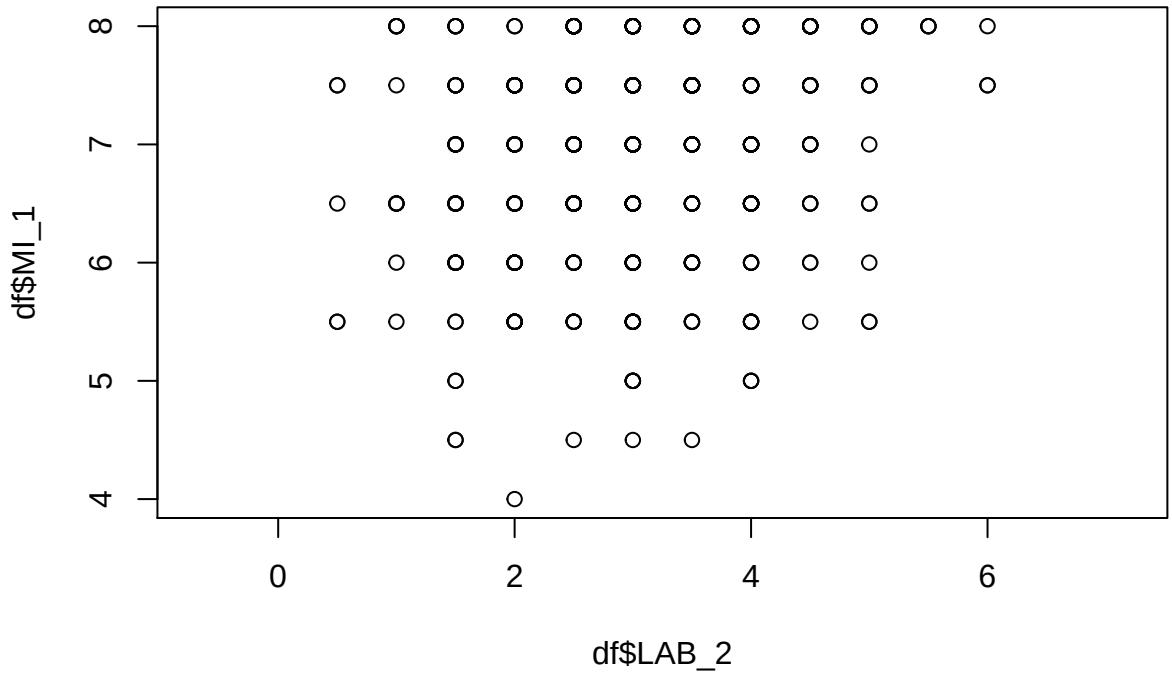

```
plot(df$LAB_1, df$MI_4, asp=1)
```



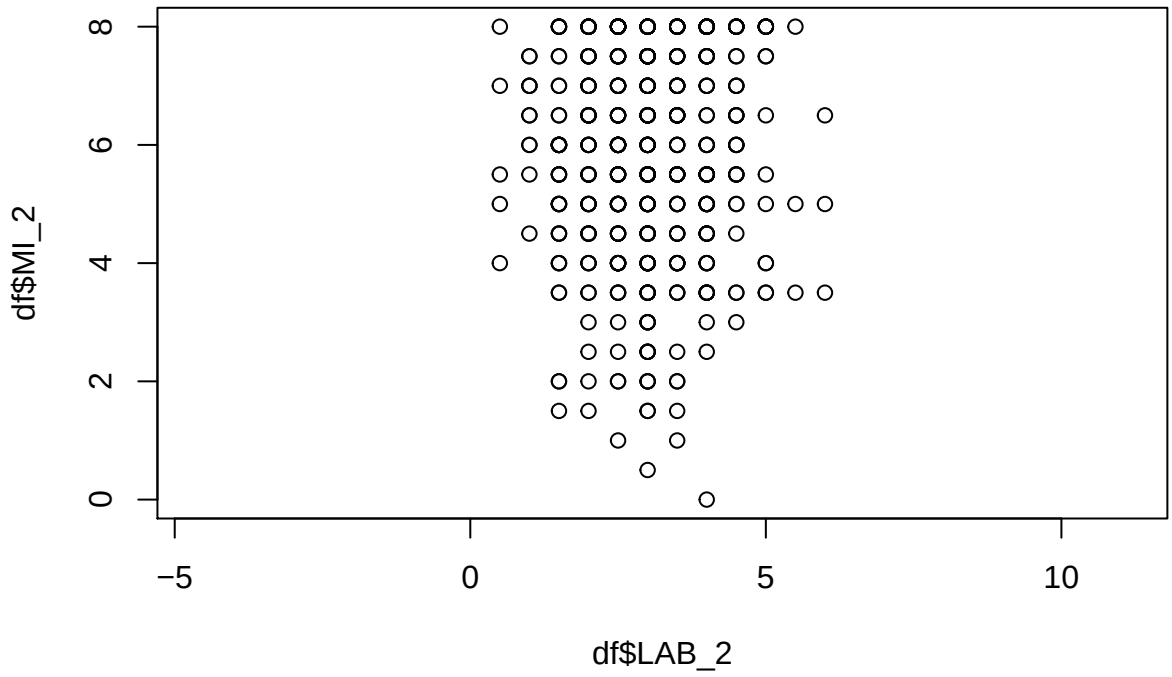

```
plot(df$LAB_1, df$MI_5, asp=1)
```



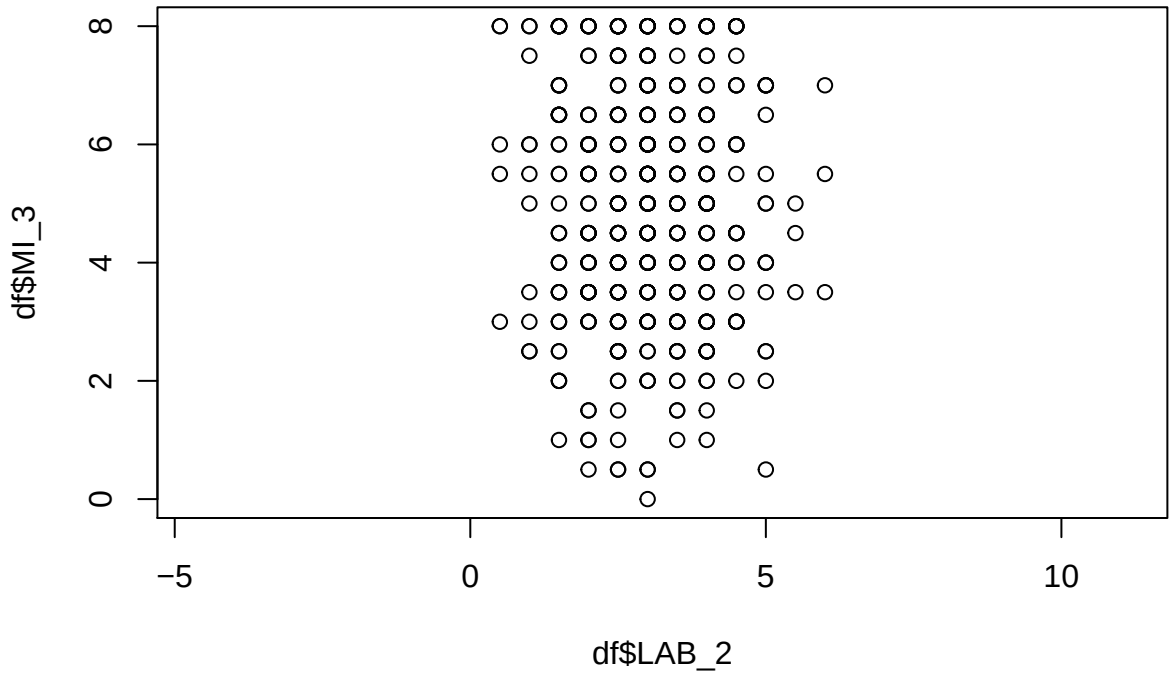

```
plot(df$LAB_2, df$MI_1, asp=1)
```



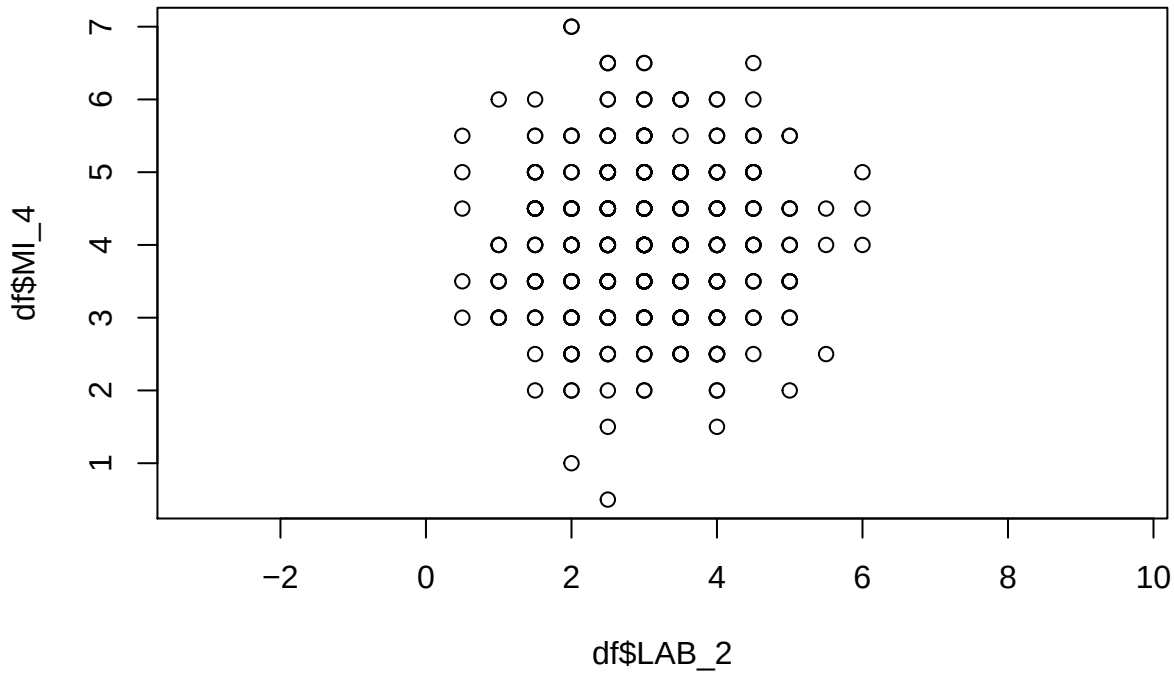

```
plot(df$LAB_2, df$MI_2, asp=1)
```



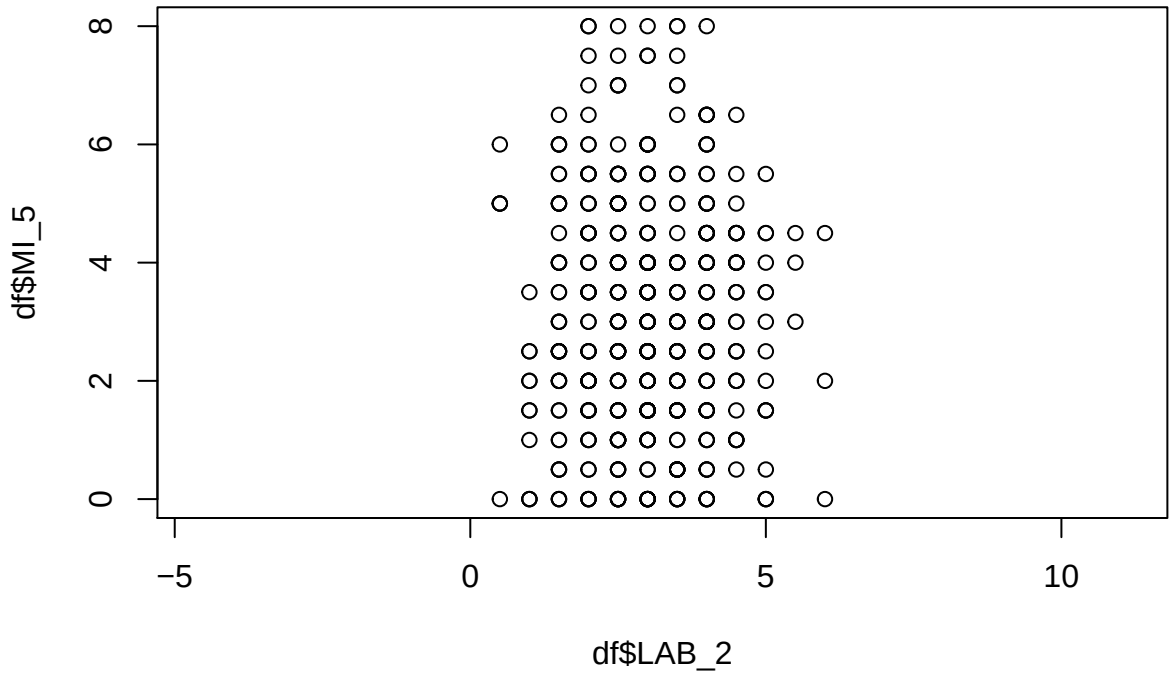

```
plot(df$LAB_2, df$MI_3, asp=1)
```



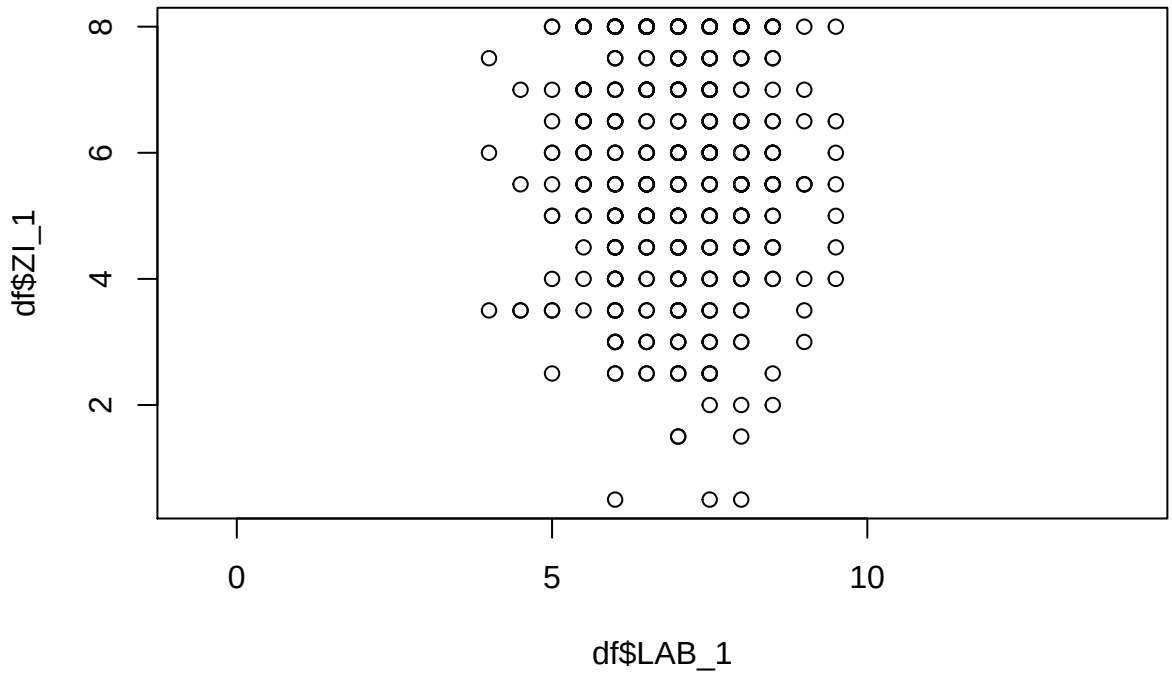

```
plot(df$LAB_2, df$MI_4, asp=1)
```



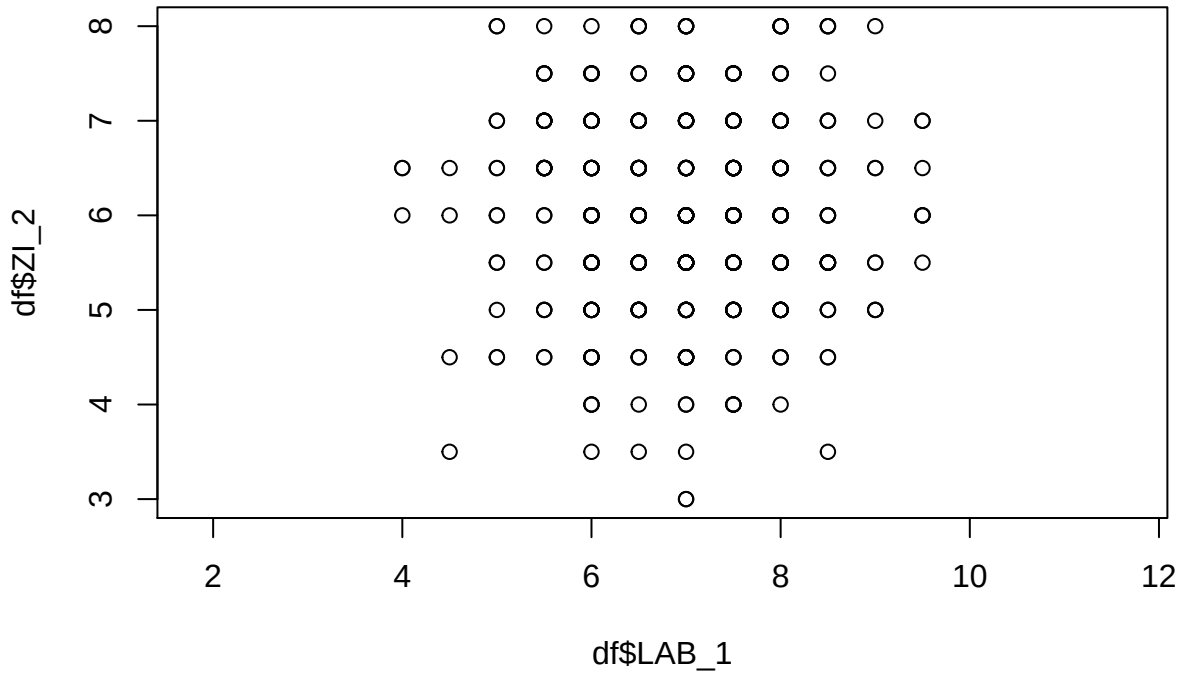

```
plot(df$LAB_2, df$MI_5, asp=1)
```



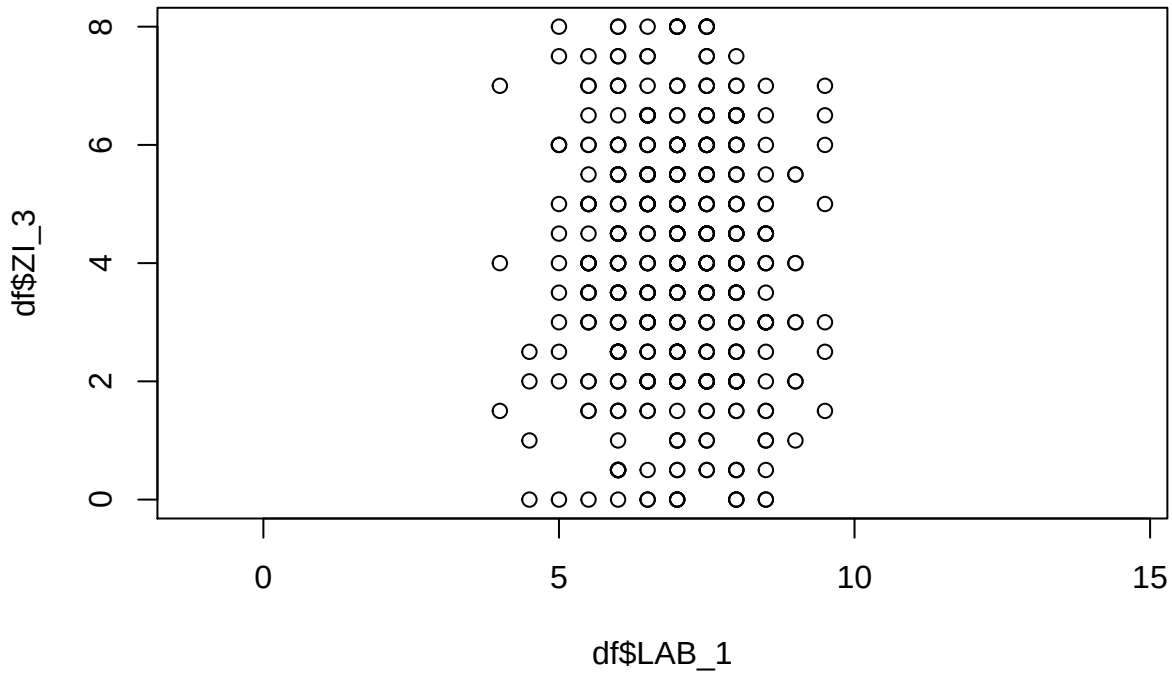

```
plot(df$LAB_1, df$ZI_1, asp=1)
```



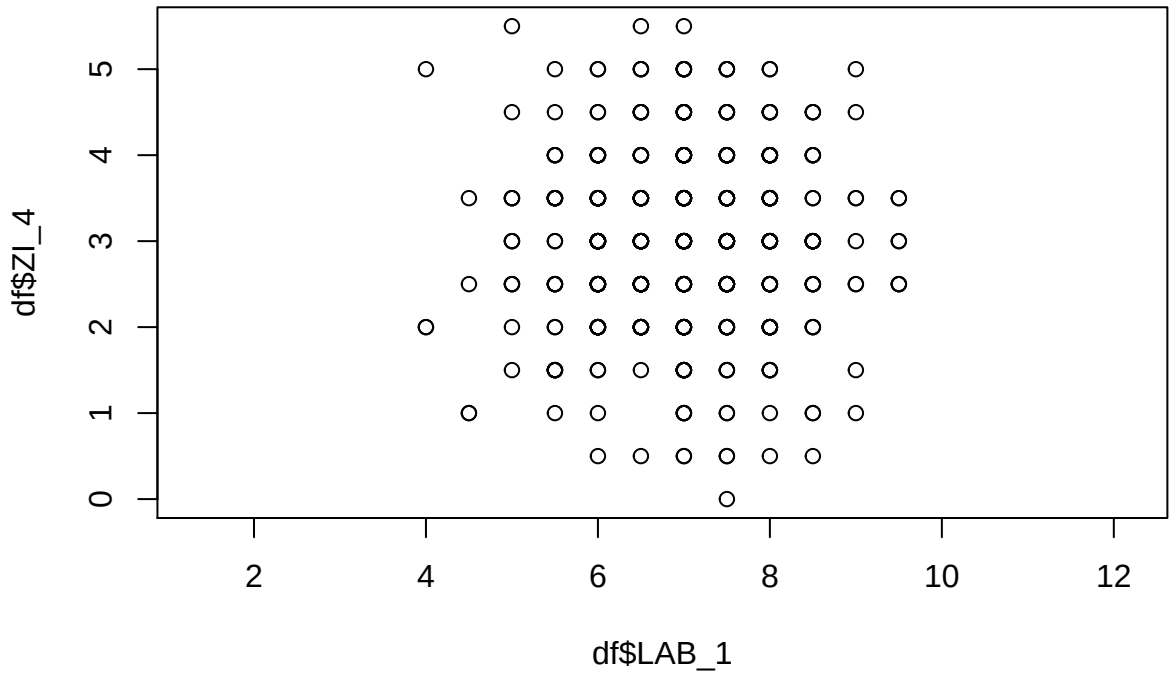

```
plot(df$LAB_1, df$ZI_2, asp=1)
```



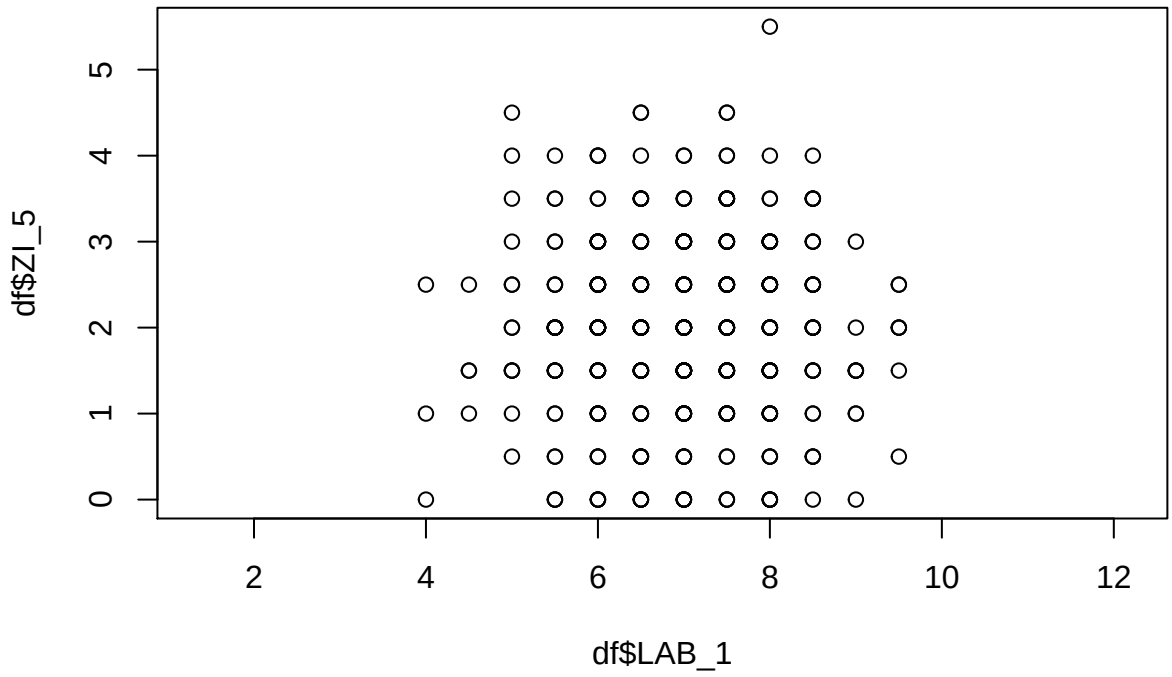

```
plot(df$LAB_1, df$ZI_3, asp=1)
```



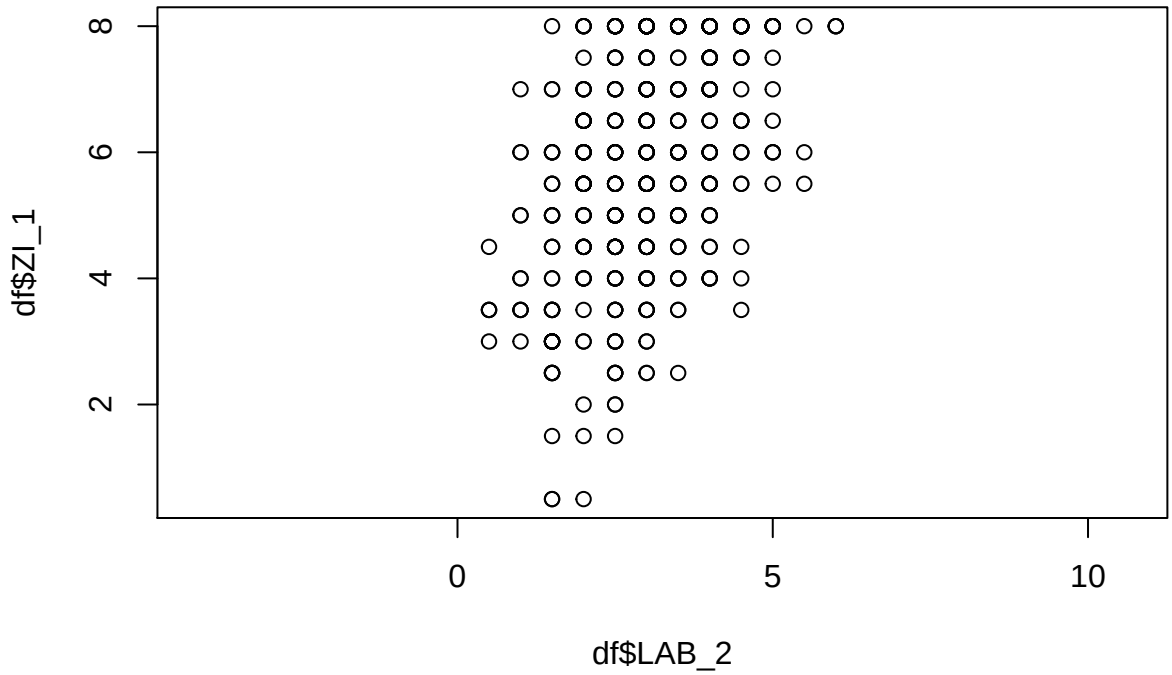

```
plot(df$LAB_1, df$ZI_4, asp=1)
```



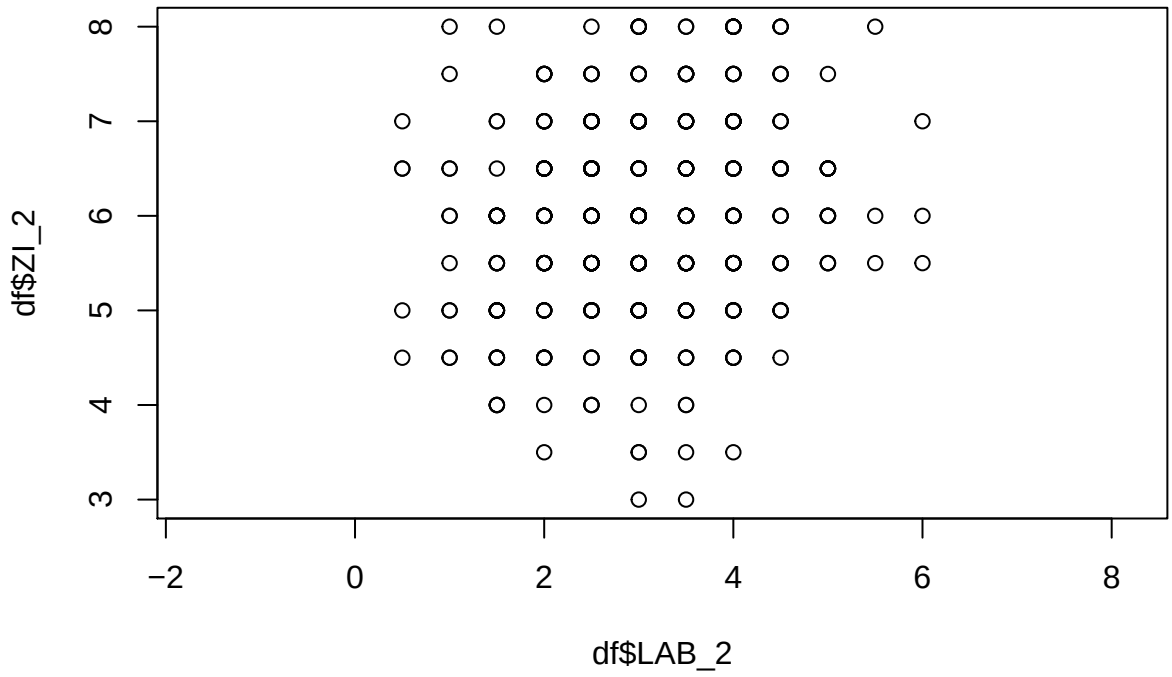

```
plot(df$LAB_1, df$ZI_5, asp=1)
```



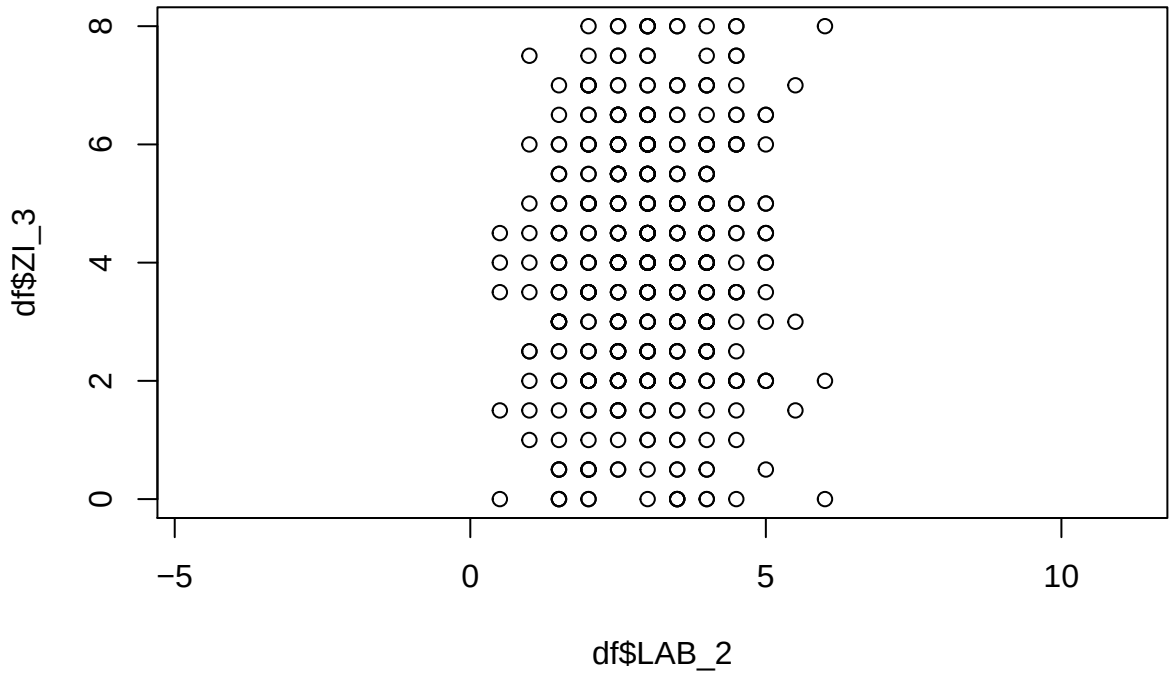

```
plot(df$LAB_2, df$ZI_1, asp=1)
```



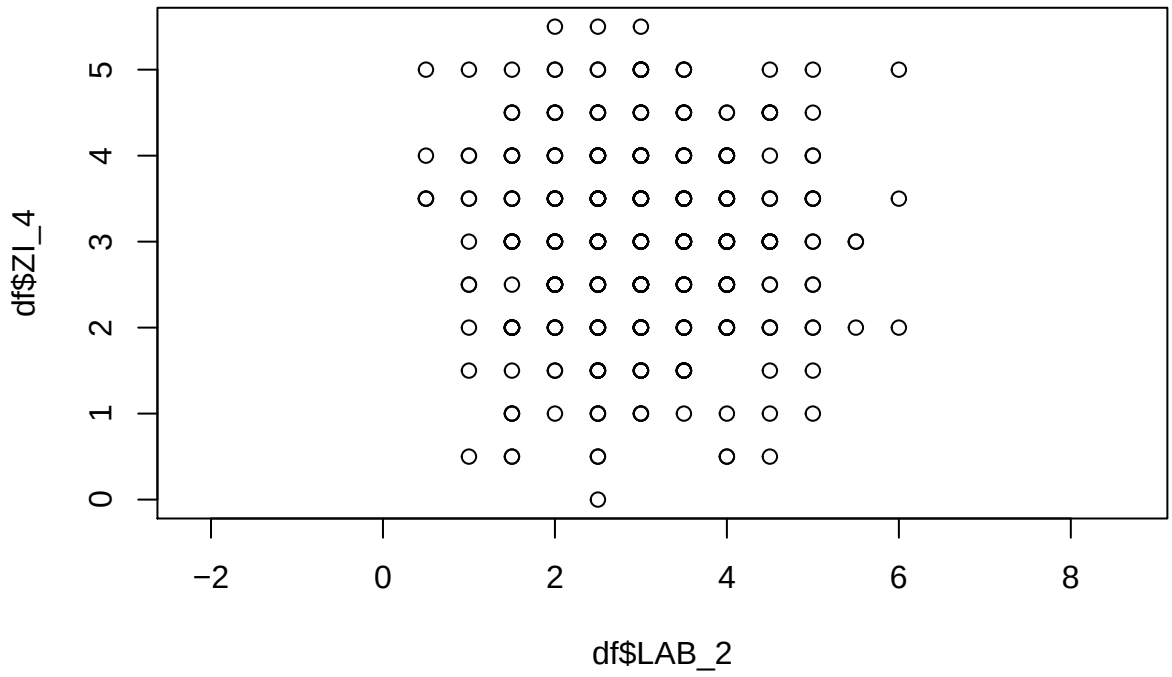

```
plot(df$LAB_2, df$ZI_2, asp=1)
```



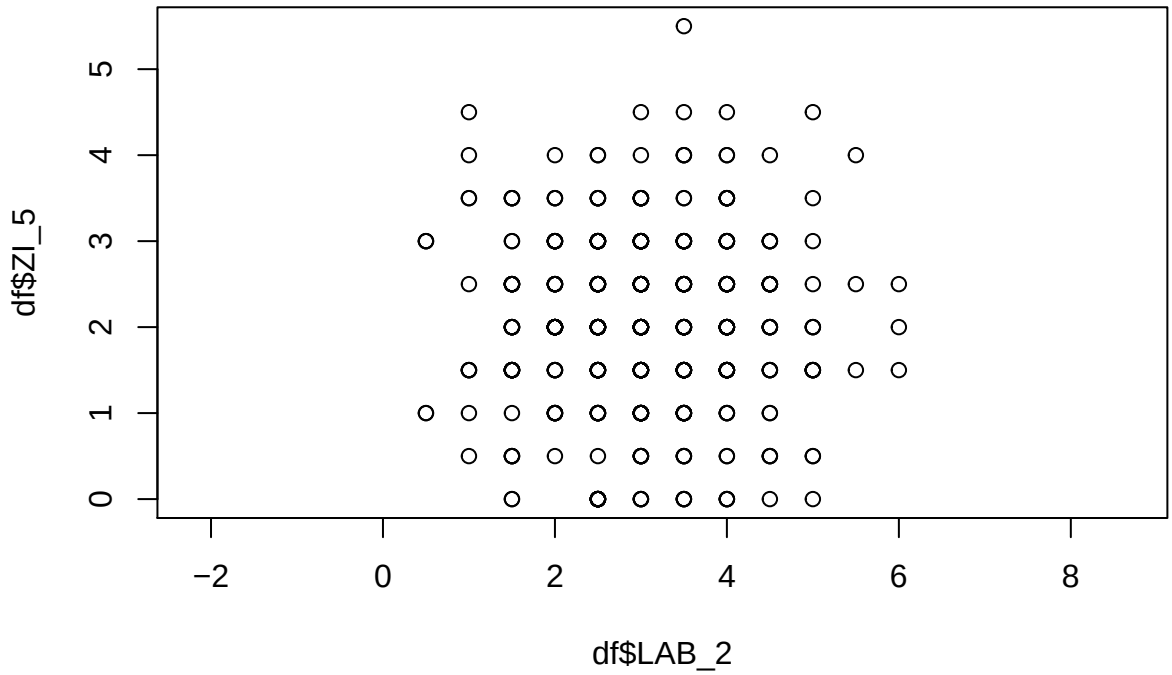

```
plot(df$LAB_2, df$ZI_3, asp=1)
```



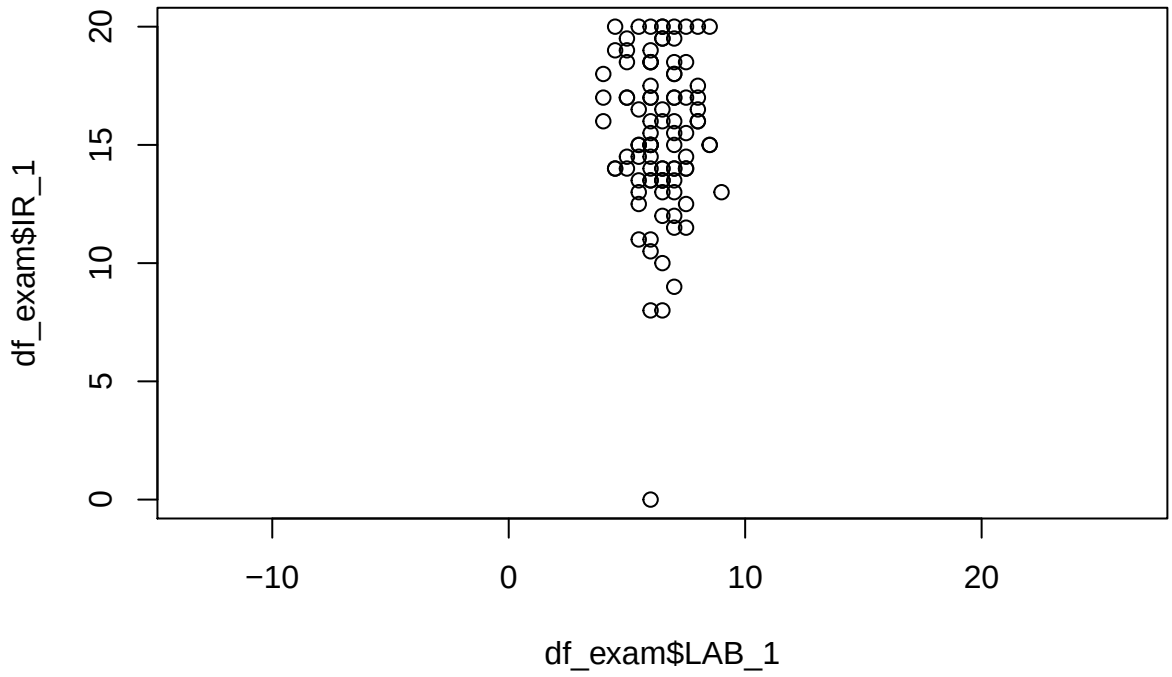

```
plot(df$LAB_2, df$ZI_4, asp=1)
```



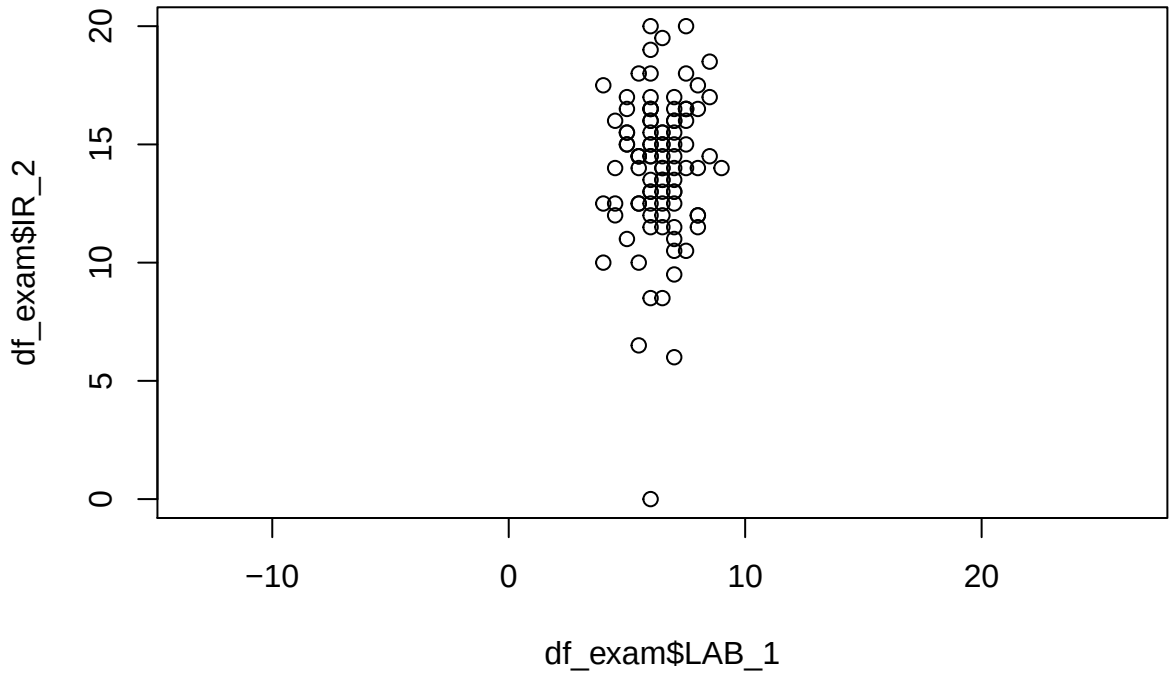

```
plot(df$LAB_2, df$ZI_5, asp=1)
```



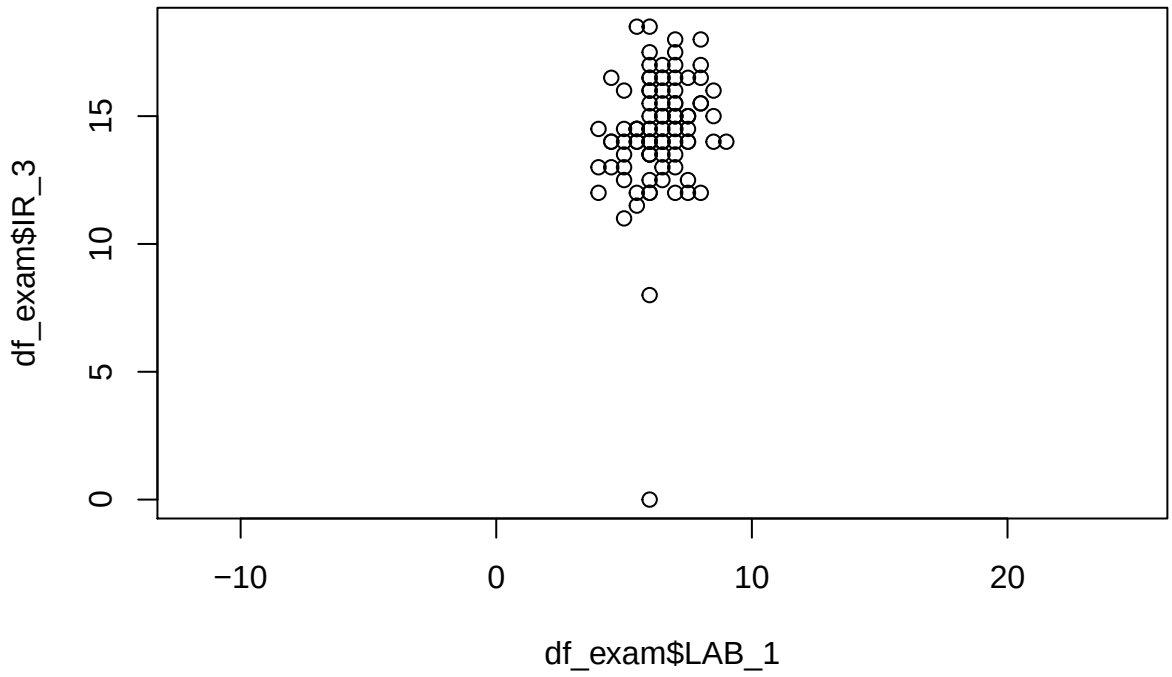

```
plot(df_exam$LAB_1, df_exam$IR_1, asp=1)
```



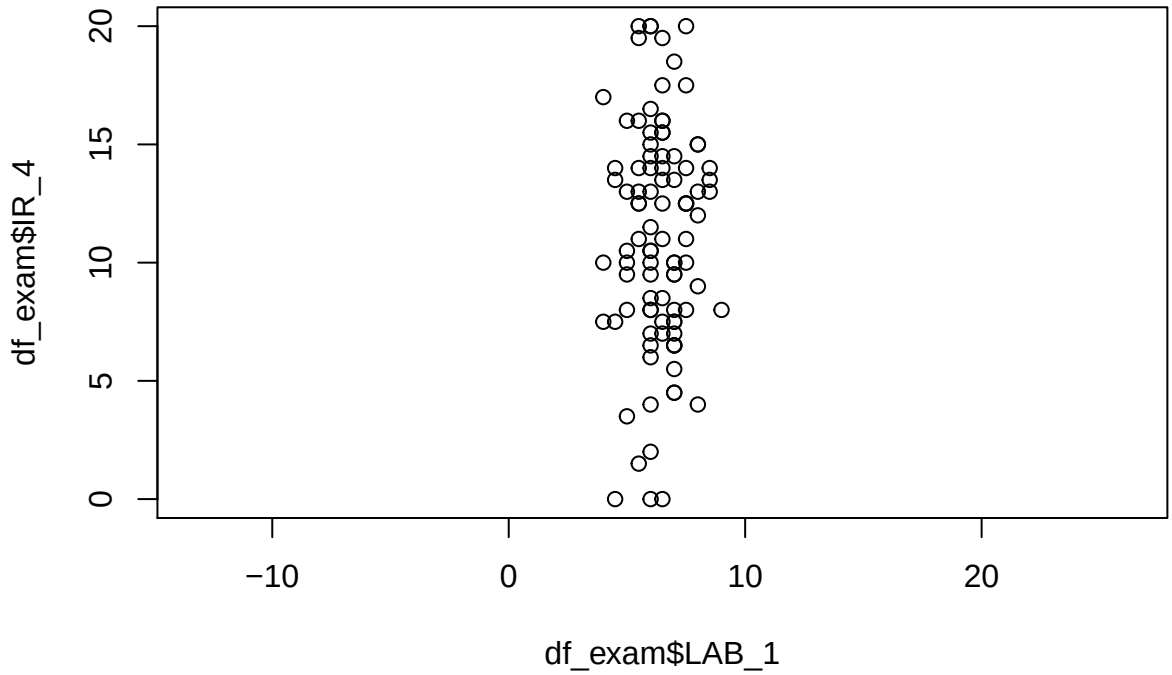

```
plot(df_exam$LAB_1, df_exam$IR_2, asp=1)
```



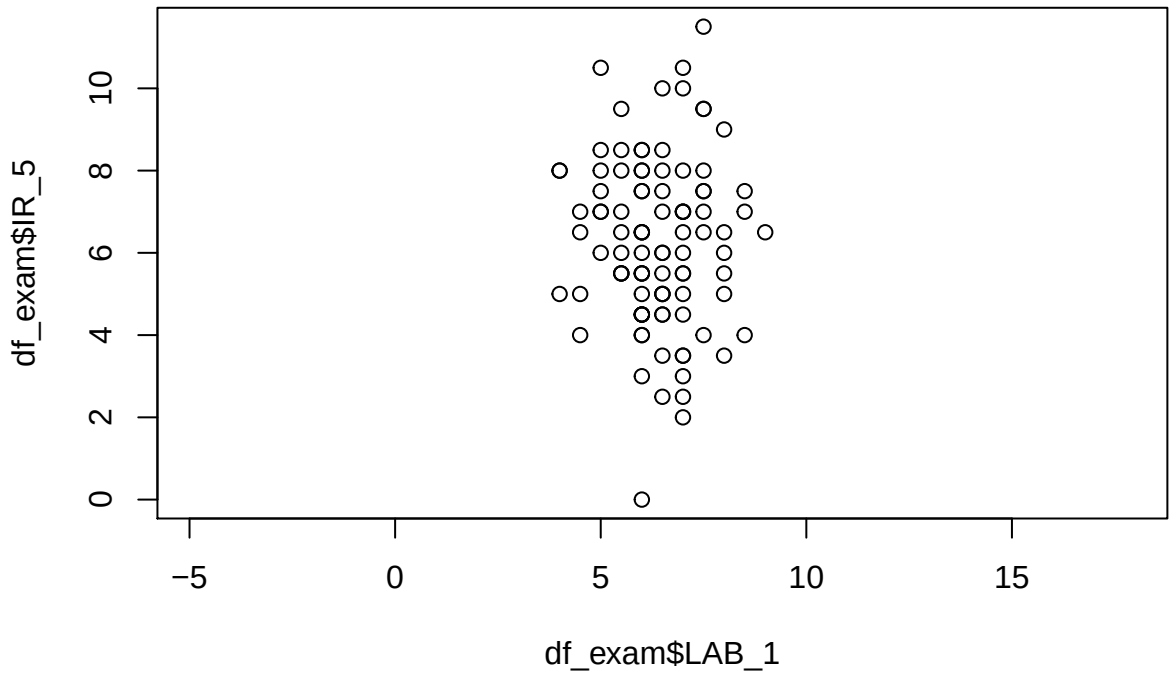

```
plot(df_exam$LAB_1, df_exam$IR_3, asp=1)
```



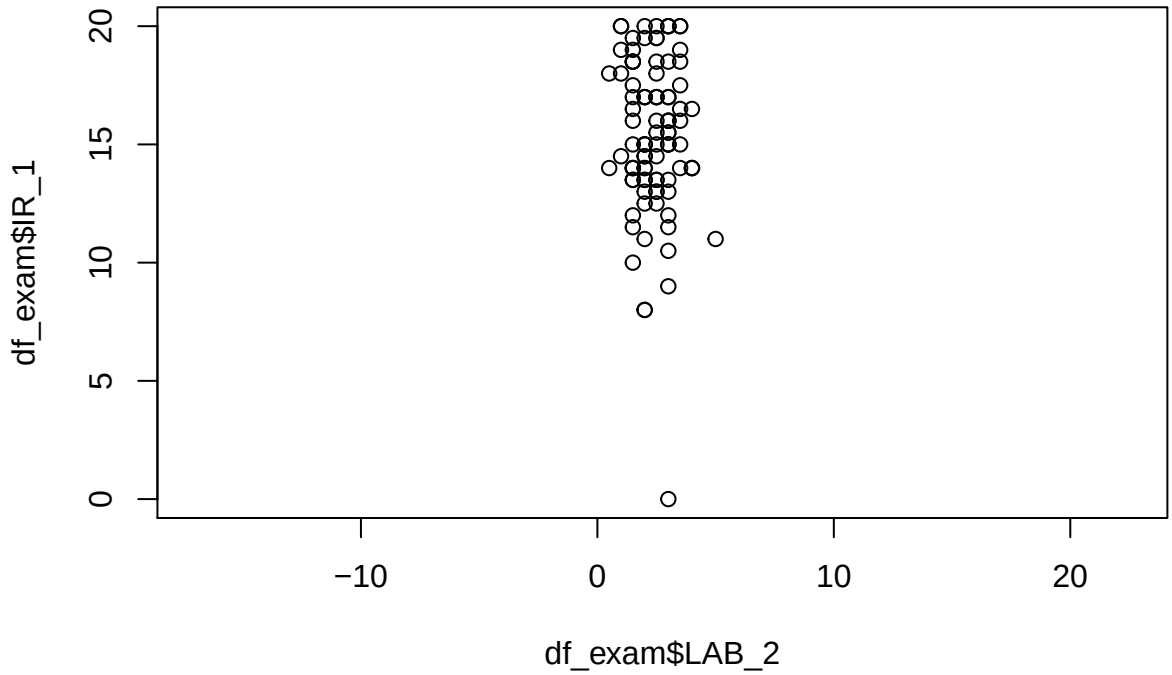

```
plot(df_exam$LAB_1, df_exam$IR_4, asp=1)
```



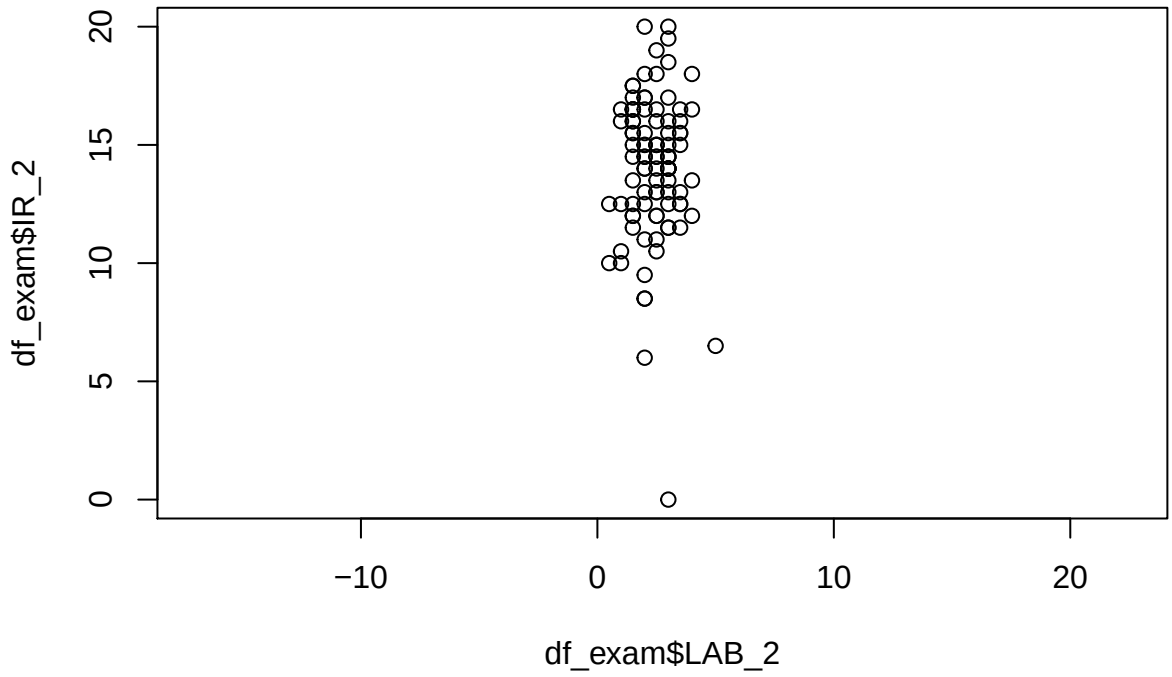

```
plot(df_exam$LAB_1, df_exam$IR_5, asp=1)
```



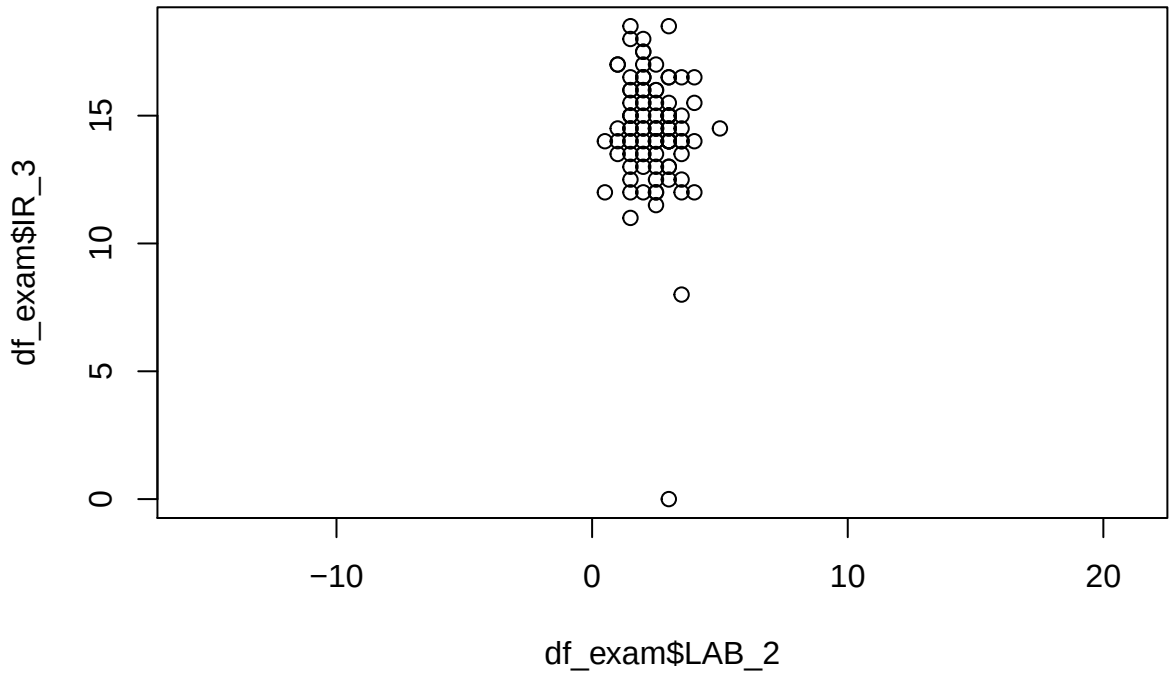

```
plot(df_exam$LAB_2, df_exam$IR_1, asp=1)
```



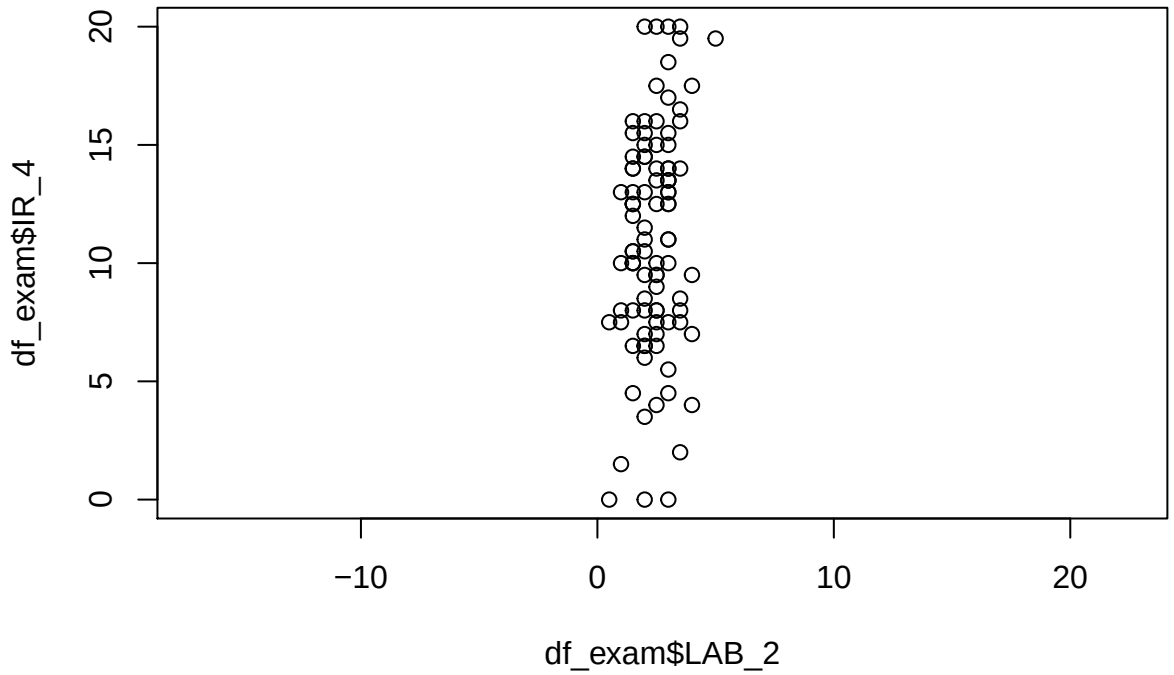

```
plot(df_exam$LAB_2, df_exam$IR_2, asp=1)
```



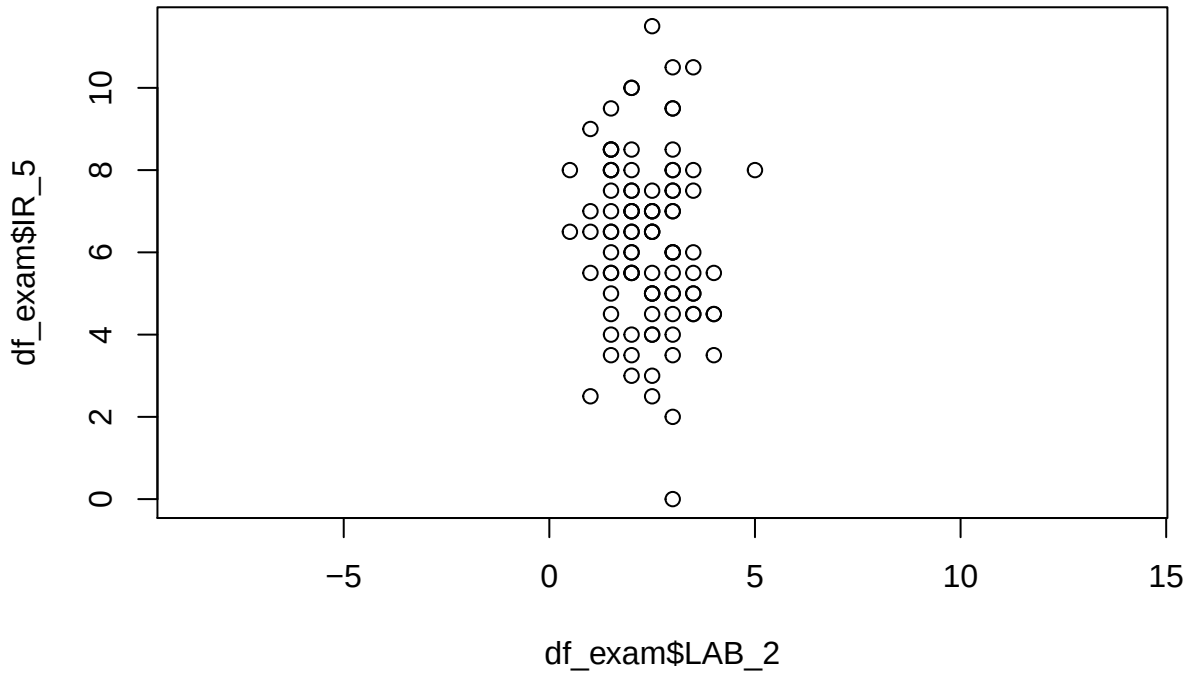

```
plot(df_exam$LAB_2, df_exam$IR_3, asp=1)
```




```
plot(df_exam$LAB_2, df_exam$IR_4, asp=1)
```




```
plot(df_exam$LAB_2, df_exam$IR_5, asp=1)
```




```
print(cor(df$LAB_1, df[c('MI_1','MI_2','MI_3','MI_4','MI_5')]))
```

```
##           MI_1      MI_2      MI_3       MI_4       MI_5
## [1,] 0.4491624 0.1298111 0.1313213 0.07703273 0.03051518
```

```
print(cor(df$LAB_2, df[c('MI_1','MI_2','MI_3','MI_4','MI_5')]))
```

```
##           MI_1       MI_2        MI_3         MI_4        MI_5
## [1,] 0.2159277 0.04211869 -0.02240775 -0.006566967 -0.01168497
```

```r
print(cor(df$LAB_1, df[c('ZI_1','ZI_2','ZI_3','ZI_4','ZI_5')]))
```

```
##              ZI_1        ZI_2        ZI_3       ZI_4       ZI_5
## [1,] -0.002025296 0.001953596 -0.01633754 0.02113404 0.01533905
```

```r
print(cor(df$LAB_2, df[c('ZI_1','ZI_2','ZI_3','ZI_4','ZI_5')]))
```

```
##           ZI_1      ZI_2       ZI_3       ZI_4        ZI_5
## [1,] 0.4705149 0.1699277 0.05961898 0.01845428 0.005028114
```

```r
print(cor(df_exam$LAB_1, df_exam[c('IR_1','IR_2','IR_3','IR_4','IR_5')]))
```

```
##             IR_1       IR_2      IR_3       IR_4        IR_5
## [1,] -0.02468903 0.09492568 0.1999346 0.04209616 -0.04005751
```

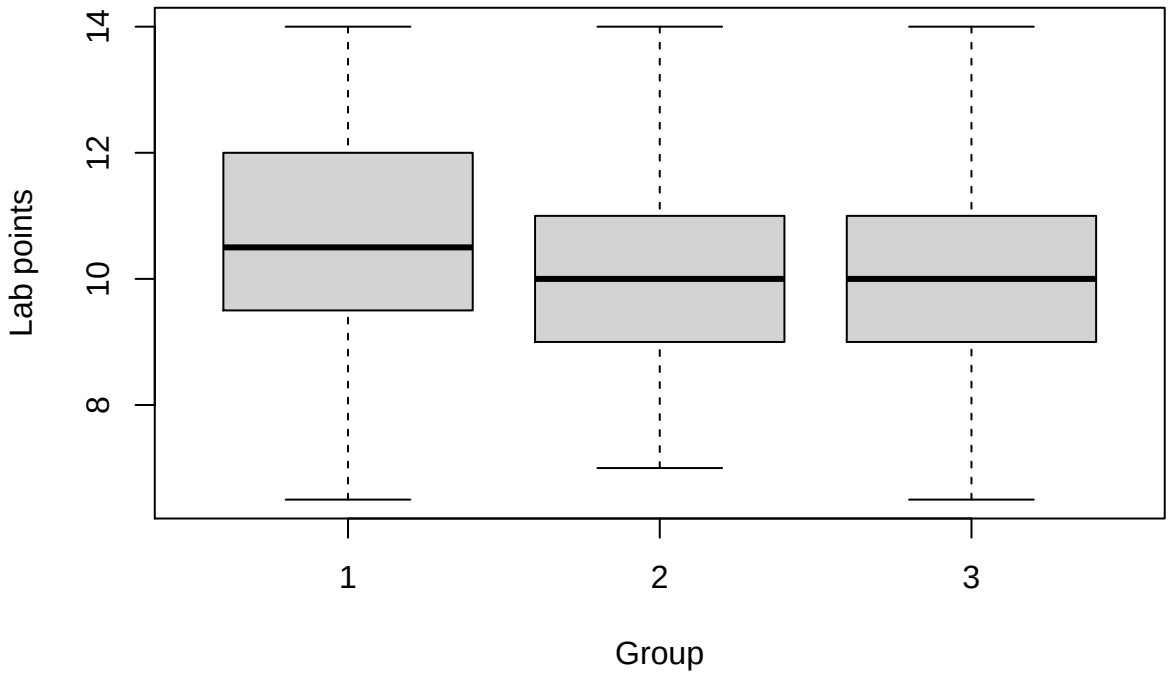
```r
print(cor(df_exam$LAB_2, df_exam[c('IR_1','IR_2','IR_3','IR_4','IR_5')]))
```

```
##             IR_1        IR_2       IR_3      IR_4       IR_5
## [1,] -0.07986937 -0.02929924 -0.1455108 0.2245866 -0.1181944
```

Većina kombinacija je slabo korelirana, jedine značajnije korelacije su kod kombinacija MI_1-LAB_1, ZI_1-LAB_2. Najviše smisla mi ima da su studenti dobili zadatak sličan tim laboratorijskim vježbama pa je otuda došla pozitivna korelacija.

Postavite i analizirajte na ovaj način još barem jedno vlastito istraživačko pitanje. * Imaju li grupe utjecaj na uspjeh na laboratorijskim vježbama kod studenata koji su prošli kontinuirano? Čini se da baš i nemaju. Vrijednosti medijana su podjednake, čak i IQR raspon je vrlo sličan pa i raspon whiskera.

```r
# Vaš kôd ovdje

# boxplot s grupama

sum_pts_lab = df_passed_cont$LAB_1 + df_passed_cont$LAB_2

boxplot(sum_pts_lab ~ df_passed_cont$Grupa,
  xlab ="Group",
  ylab = "Lab points",
)
```



4.2. Regresijska analiza

Razmotrimo u kakvom su odnosu zadatci ispitnog roka s ostalim aktivnostima iz kontinuirane nastave. Istražite odnos koristeći model multivarijatne linearne regresije. Procijenite model gdje su zavisne varijable bodovi zadataka s ispitnog roka, odaberite konačni skup ulaznih varijabli i provjerite adekvatnost modela.

```
# Vaš kôd ovdje

reg <- lm(cbind(IR_1, IR_2, IR_3, IR_4, IR_5) ~ ., data=df)

summary(reg)
```

```
## Response IR_1 :
##
## Call:
## lm(formula = IR_1 ~ MI_1 + MI_2 + MI_3 + MI_4 + MI_5 + LAB_1 +
##     ZI_1 + ZI_2 + ZI_3 + ZI_4 + ZI_5 + LAB_2 + Grupa + take_exam +
##     mi + zi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6281  -0.3305  -0.0234   0.3006   4.6425
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.3039700  1.0452655  -2.204  0.02798 *
## MI_1         0.4235683  0.0941798   4.497 8.63e-06 ***
## MI_2        -0.0928397  0.0431311  -2.153  0.03186 *
## MI_3         0.0900438  0.0387972   2.321  0.02071 *
## MI_4         0.0292270  0.0670032   0.436  0.66289
## MI_5        -0.0261626  0.0360925  -0.725  0.46888
## LAB_1       -0.1488992  0.0773629  -1.925  0.05486 .
## ZI_1        -0.0515199  0.0508841  -1.012  0.31181
```

```
## ZI_2         0.0632019  0.0736198   0.858  0.39105
## ZI_3        -0.0422448  0.0379693  -1.113  0.26644
## ZI_4         0.0205913  0.0671059   0.307  0.75909
## ZI_5         0.0098559  0.0658256   0.150  0.88104
## LAB_2        0.0005544  0.0804370   0.007  0.99450
## Grupa        0.2380993  0.0852490   2.793  0.00543 **
## take_exam   15.4295914  0.2381366  64.793  < 2e-16 ***
## mi                 NA         NA      NA       NA
## zi                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.411 on 480 degrees of freedom
## Multiple R-squared:  0.9508, Adjusted R-squared:  0.9493
## F-statistic: 662.1 on 14 and 480 DF,  p-value: < 2.2e-16
##
##
## Response IR_2 :
##
## Call:
## lm(formula = IR_2 ~ MI_1 + MI_2 + MI_3 + MI_4 + MI_5 + LAB_1 +
##     ZI_1 + ZI_2 + ZI_3 + ZI_4 + ZI_5 + LAB_2 + Grupa + take_exam +
##     mi + zi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3854  -0.2628   0.0107   0.2744   5.4669
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.170616   0.982568  -1.191 0.234092
## MI_1        -0.211775   0.088531  -2.392 0.017136 *
## MI_2         0.029378   0.040544   0.725 0.469046
## MI_3        -0.025852   0.036470  -0.709 0.478759
## MI_4         0.227098   0.062984   3.606 0.000344 ***
## MI_5         0.072820   0.033928   2.146 0.032346 *
## LAB_1        0.157685   0.072723   2.168 0.030626 *
## ZI_1         0.007807   0.047832   0.163 0.870417
## ZI_2        -0.022571   0.069204  -0.326 0.744447
## ZI_3         0.014411   0.035692   0.404 0.686569
## ZI_4        -0.031880   0.063081  -0.505 0.613523
## ZI_5         0.056967   0.061877   0.921 0.357702
## LAB_2        0.004289   0.075612   0.057 0.954790
## Grupa        0.166753   0.080136   2.081 0.037974 *
## take_exam   14.225818   0.223853  63.550  < 2e-16 ***
## mi                 NA         NA      NA       NA
## zi                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.326 on 480 degrees of freedom
## Multiple R-squared:  0.9482, Adjusted R-squared:  0.9467
## F-statistic: 627.4 on 14 and 480 DF,  p-value: < 2.2e-16
##
```

```
##
## Response IR_3 :
##
## Call:
## lm(formula = IR_3 ~ MI_1 + MI_2 + MI_3 + MI_4 + MI_5 + LAB_1 +
##     ZI_1 + ZI_2 + ZI_3 + ZI_4 + ZI_5 + LAB_2 + Grupa + take_exam +
##     mi + zi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7392  -0.2366   0.0214   0.2604   3.7245
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.72177    0.72284  -0.999   0.3185
## MI_1        -0.02138    0.06513  -0.328   0.7429
## MI_2         0.03734    0.02983   1.252   0.2112
## MI_3        -0.11380    0.02683  -4.241 2.67e-05 ***
## MI_4         0.01663    0.04634   0.359   0.7198
## MI_5         0.02888    0.02496   1.157   0.2479
## LAB_1        0.21646    0.05350   4.046 6.07e-05 ***
## ZI_1         0.18839    0.03519   5.354 1.34e-07 ***
## ZI_2        -0.12435    0.05091  -2.442   0.0149 *
## ZI_3         0.01080    0.02626   0.411   0.6811
## ZI_4        -0.10009    0.04641  -2.157   0.0315 *
## ZI_5         0.02265    0.04552   0.498   0.6190
## LAB_2       -0.24019    0.05563  -4.318 1.91e-05 ***
## Grupa        0.05466    0.05895   0.927   0.3543
## take_exam   14.50801    0.16468  88.098  < 2e-16 ***
## mi                NA         NA      NA       NA
## zi                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9755 on 480 degrees of freedom
## Multiple R-squared:  0.9725, Adjusted R-squared:  0.9717
## F-statistic:  1214 on 14 and 480 DF,  p-value: < 2.2e-16
##
##
## Response IR_4 :
##
## Call:
## lm(formula = IR_4 ~ MI_1 + MI_2 + MI_3 + MI_4 + MI_5 + LAB_1 +
##     ZI_1 + ZI_2 + ZI_3 + ZI_4 + ZI_5 + LAB_2 + Grupa + take_exam +
##     mi + zi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1221  -0.4142  -0.0169   0.4364   9.5025
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.393637   1.514871  -1.580   0.1147
## MI_1         0.019554   0.136492   0.143   0.8861
```

```
## MI_2        -0.153238   0.062509  -2.451   0.0146 *
## MI_3         0.083052   0.056228   1.477   0.1403
## MI_4         0.105260   0.097106   1.084   0.2789
## MI_5        -0.016956   0.052308  -0.324   0.7460
## LAB_1       -0.004795   0.112120  -0.043   0.9659
## ZI_1        -0.119986   0.073745  -1.627   0.1044
## ZI_2         0.443838   0.106695   4.160 3.77e-05 ***
## ZI_3         0.019973   0.055028   0.363   0.7168
## ZI_4        -0.132862   0.097254  -1.366   0.1725
## ZI_5        -0.029644   0.095399  -0.311   0.7561
## LAB_2        0.234141   0.116575   2.009   0.0451 *
## Grupa        0.042960   0.123549   0.348   0.7282
## take_exam   11.252984   0.345124  32.606  < 2e-16 ***
## mi                NA         NA      NA       NA
## zi                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.044 on 480 degrees of freedom
## Multiple R-squared:  0.8274, Adjusted R-squared:  0.8224
## F-statistic: 164.4 on 14 and 480 DF,  p-value: < 2.2e-16
##
##
## Response IR_5 :
##
## Call:
## lm(formula = IR_5 ~ MI_1 + MI_2 + MI_3 + MI_4 + MI_5 + LAB_1 +
##     ZI_1 + ZI_2 + ZI_3 + ZI_4 + ZI_5 + LAB_2 + Grupa + take_exam +
##     mi + zi, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0608 -0.1821  0.0065  0.1747  4.8870
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.045363   0.661005  -1.581    0.114
## MI_1        -0.030737   0.059557  -0.516    0.606
## MI_2         0.040143   0.027275   1.472    0.142
## MI_3        -0.011383   0.024535  -0.464    0.643
## MI_4         0.015266   0.042371   0.360    0.719
## MI_5         0.020092   0.022824   0.880    0.379
## LAB_1        0.021769   0.048923   0.445    0.657
## ZI_1         0.006414   0.032178   0.199    0.842
## ZI_2         0.029567   0.046556   0.635    0.526
## ZI_3         0.005016   0.024011   0.209    0.835
## ZI_4         0.036733   0.042436   0.866    0.387
## ZI_5         0.189265   0.041627   4.547 6.91e-06 ***
## LAB_2       -0.027703   0.050867  -0.545    0.586
## Grupa        0.064768   0.053910   1.201    0.230
## take_exam    6.436060   0.150593  42.738  < 2e-16 ***
## mi                NA         NA      NA       NA
## zi                NA         NA      NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8921 on 480 degrees of freedom
## Multiple R-squared:  0.8888, Adjusted R-squared:  0.8856
## F-statistic: 274.1 on 14 and 480 DF,  p-value: < 2.2e-16
```