Иван Адамов

Факултетен номер 71534

Специалност Информационни системи, 2-ра административна група

1) Прочетете данните и ги запишете в data frame в R;

```
> filename = "pokemon.csv"
> pokemonData = read.csv(filename, header=TRUE)
> class(pokemonData)
[1] "data.frame"
>
```

2) Генерирайте си подизвадка от 600 наблюдения. За целта нека f_nr е вашият факултетен номер. Задайте състояние на генератора на случайни числа в R чрез set.seed(f_nr). С помощта на подходяща функция генерирайте извадка без връщане на числата от 1 до 705 като не забравяте да я запишете във вектор. Използвайте вектора, за да зашишете само редовете със съответните индекси в нов дейтафрейм и работете с него оттук нататък. Изкарайте на екрана първите няколко (5-6) наблюдения;

```
>
> f_nr = 71534
> set.seed(f_nr)
>
> indexes = sample(1:705, 600, replace=F)
> subsample = subset(pokemonData, pokemonData$Number %in% indexes)
> head(subsample)
  Number       Name Type1  Type2 Attack Defense Height Weight
1      1  Bulbasaur Grass Poison     49      49   0.71    6.9
3      3   Venusaur Grass Poison     82      83   2.01  100.0
4      4 Charmander  Fire             52      43   0.61    8.5
5      5 Charmeleon  Fire             64      58   1.09   19.0
6      6  Charizard  Fire Flying     84      78   1.70   90.5
7      7   Squirtle Water            48      65   0.51    9.0
> nrow(subsample)
[1] 600
>
```

3) Какъв вид данни (качествени/количествени, непрекъснати/дискретни) са записани във всяка от променливите?

```
>
> str(subsample)
'data.frame':   600 obs. of  8 variables:
 $ Number : int  1 3 4 5 6 7 8 9 10 11 ...
 $ Name   : Factor w/ 705 levels "Abomasnow","Abra",..: 67 656 86 87 85 583 674 53 81 386 ...
 $ Type1  : Factor w/ 18 levels "Bug","Dark","Dragon",..: 10 10 7 7 7 18 18 18 1 1 ...
 $ Type2  : Factor w/ 19 levels "","Bug","Dark",..: 15 15 1 1 9 1 1 1 1 1 ...
 $ Attack : int  49 82 52 64 84 48 63 83 30 20 ...
 $ Defense: int  49 83 43 58 78 65 80 100 35 55 ...
 $ Height : num  0.71 2.01 0.61 1.09 1.7 0.51 0.99 1.6 0.3 0.71 ...
 $ Weight : num  6.9 100 8.5 19 90.5 9 22.5 85.5 2.9 9.9 ...
>
```

4) Изведете дескриптивни статистики за всяка една от променливите;

Количествени са Attack, Deffense, Height и Weight, а останалите са качествени. Type1 и Type2 са непрекъснати, а останалите са дискретни.

```
>
> summary(subsample)
     Number                Name         Type1          Type2         Attack
 Min.   :  1.0   Abomasnow :  1   Water  : 93          :312   Min.   :  5.00
 1st Qu.:176.2   Absol     :  1   Normal : 81   Flying : 74   1st Qu.: 53.00
 Median :352.5   Accelgor  :  1   Bug    : 54   Poison : 27   Median : 72.00
 Mean   :353.0   Aerodactyl:  1   Grass  : 50   Ground : 25   Mean   : 74.12
 3rd Qu.:532.2   Aggron    :  1   Fire   : 40   Psychic: 21   3rd Qu.: 93.25
 Max.   :705.0   Aipom     :  1   Psychic: 40   Fairy  : 16   Max.   :165.00
                 (Other)   :594   (Other):242   (Other):125
    Defense          Height           Weight
 Min.   :  5.0   Min.   : 0.100   Min.   :  0.10
 1st Qu.: 50.0   1st Qu.: 0.610   1st Qu.:  9.50
 Median : 65.0   Median : 0.990   Median : 28.00
 Mean   : 70.2   Mean   : 1.101   Mean   : 53.45
 3rd Qu.: 85.0   3rd Qu.: 1.400   3rd Qu.: 59.70
 Max.   :230.0   Max.   :14.500   Max.   :683.00
```

5) Изведете редовете на най-високия и на най-лекия покемон;

```
>
> subsample[which.min(subsample$Weight),]
   Number   Name Type1  Type2 Attack Defense Height Weight
92     92 Gastly Ghost Poison     35      30    1.3    0.1
> subsample[which.max(subsample$Height),]
    Number    Name Type1 Type2 Attack Defense Height Weight
321    321 Wailord Water            90      45   14.5    398
> s
```

6) Изведете редовете на покемоните с общ брой точки за атака и защита над 220;

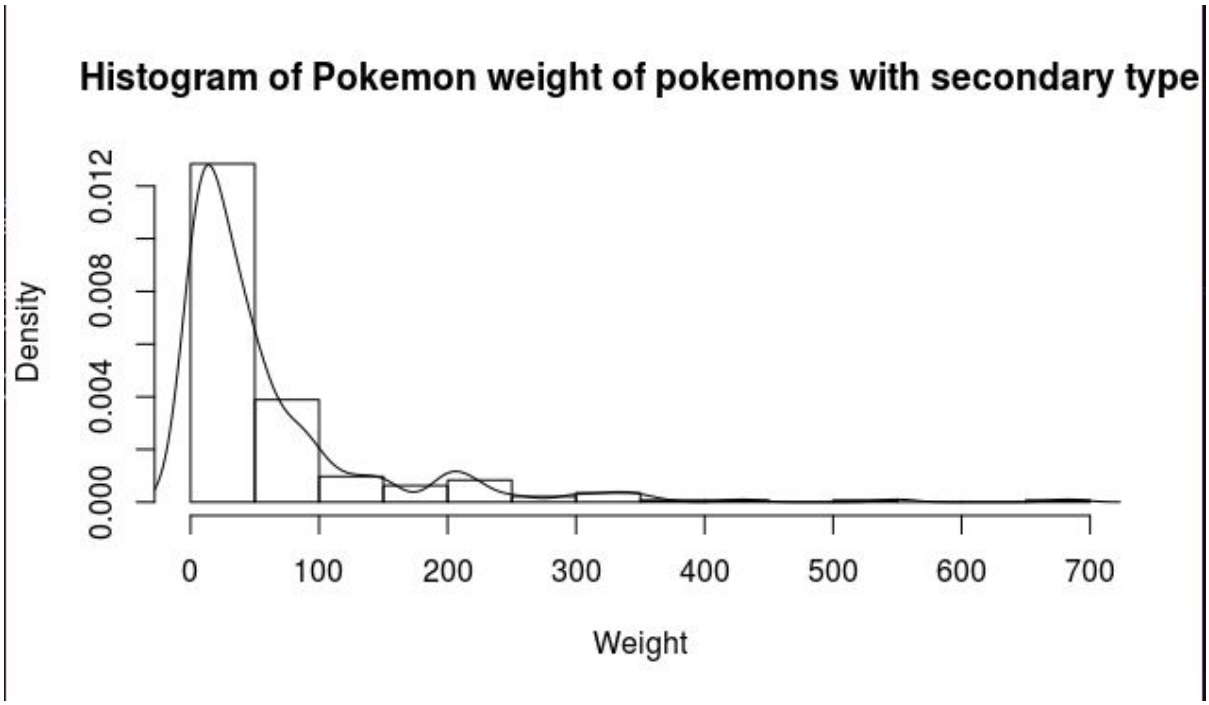
```
>
> nrow(subsample[subsample$Attack+subsample$Defense > 220,])
[1] 39
>
```

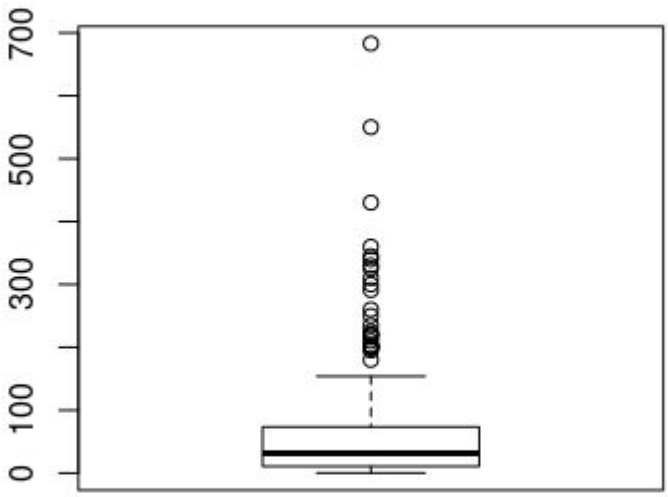
7) Колко на брой покемони имат първичен или вторичен тип "Dragon" или "Flying" и са високи над един метър?

```
>
> types = c("Dragon", "Flying")
> nrow(subsample[(subsample$Height > 1 & (subsample$Type1 %in% types | subsample$Type2 %in% types)),])
[1] 58
>
```

8) Направете хистограма на теглото само на покемоните с вторичен тип и нанесете графика на плътността върху нея. Симетрично ли са разположени данните?

```
>
> secondaries = subset(subsample, !subsample$Type2  == "", select = c("Weight"))
> vectorizaedSecondaries = (unlist(secondaries))
>
> hist(vectorizaedSecondaries, main="Histogram of Pokemon weight of pokemons with secondary type", xlab="Weight", probability=TRUE)
> lines(density(vectorizaedSecondaries))
>
> boxplot(vectorizaedSecondaries)
>
```

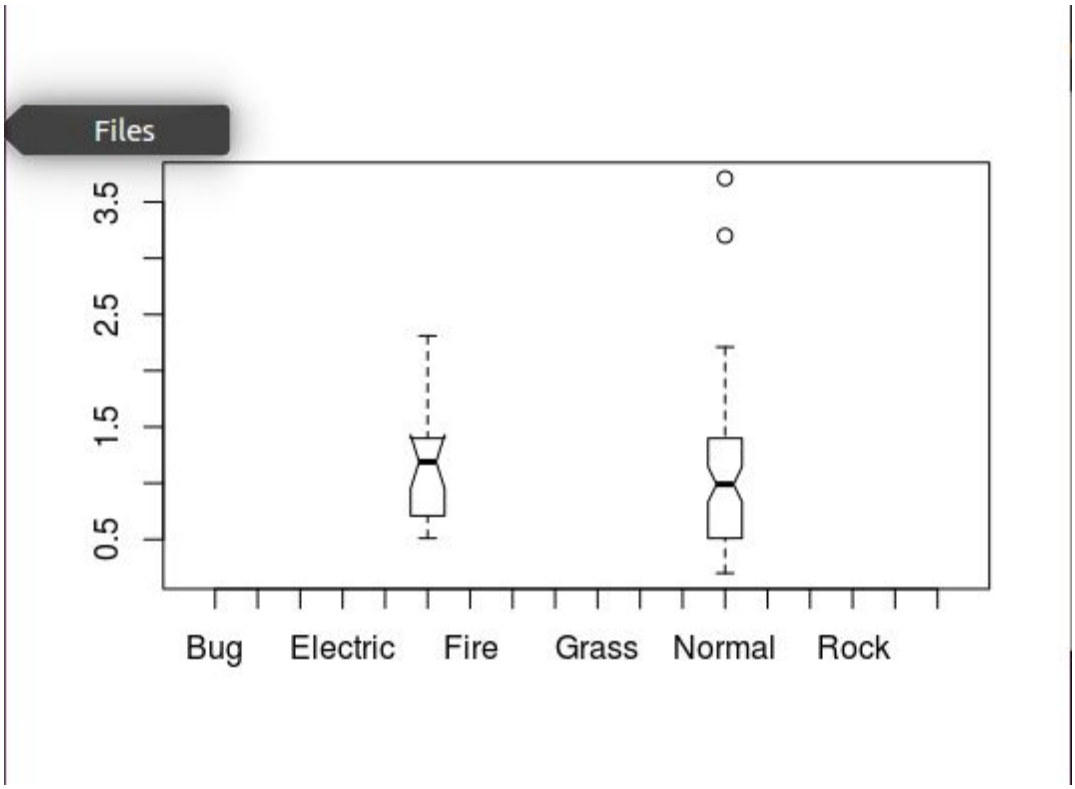




Данните не са симетрични.

9) За покемоните с първичен тип "Normal" или "Fighting" изследвайте съвместно променливите Type1 и Height с подходящ графичен метод. Забелязвате ли outlier-и? Сравнете извадковите средни и медианите в двете групи и направете извод;

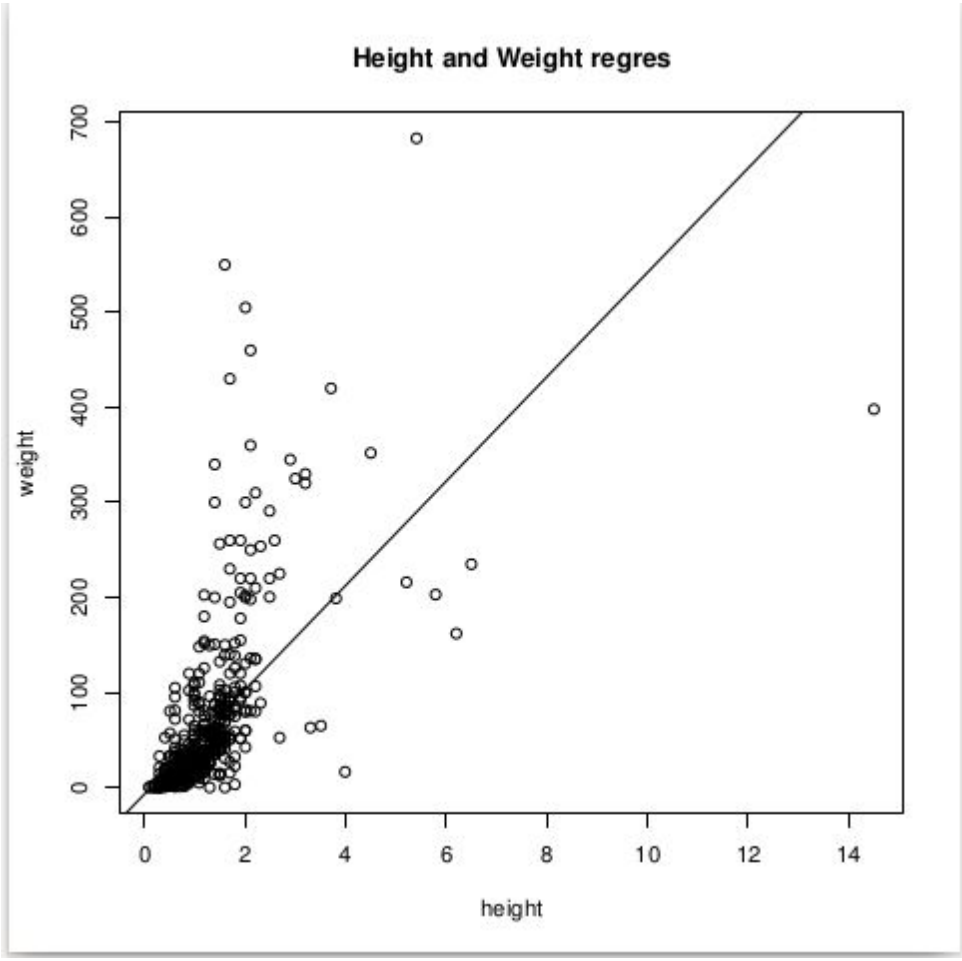
```
>
> primaries = subset(subsample, subsample$Type1  %in% c("Normal", "Fighting"), select = c("Type1", "Height"))
>
> box = boxplot(primaries$Height ~ primaries$Type1, notch=TRUE)
Warning message:
In bxp(list(stats = c(NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,  :
  some notches went outside hinges ('box'): maybe set notch=FALSE
> print("Outliers: ")
[1] "Outliers: "
> print(box$out)
[1] 3.71 3.20
>
> aggregate(primaries$Height ~ primaries$Type1, data=primaries, FUN="mean")
  primaries$Type1 primaries$Height
1        Fighting         1.149048
2          Normal         1.019012
> aggregate(primaries$Height ~ primaries$Type1, data=primaries, FUN="median")
  primaries$Type1 primaries$Height
1        Fighting             1.19
2          Normal             0.99
```



Виждат се и outlier-ите и това, че Fighter-ите са по-високи от Normal-ите.

10) Изследвайте съвместно променливите Height и Weight с подходящ графичен метод. Бихте ли казали, че съществува линейна връзка между тях? Намерете корелацията между величините и коментирайте стойността `и. Начертайте регресионна права (линейната функция, която най-добре приближава функционалната зависимост). Ако е наблюдаван нов вид покемон с височина 2.1 метра, какво се очаква да е теглото му на базата на линейния модел?

```
> height = subsample$Height
> weight = subsample$Weight
>
> plot(height, weight, main="Height and Weight regres")
>
> relation = lm(weight ~ height)
> relation

Call:
lm(formula = weight ~ height)

Coefficients:
(Intercept)       height
     -6.962       54.888

> summary(relation)

Call:
lm(formula = weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max
-390.91  -21.05  -11.94   -2.02  469.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.962      3.858  -1.804   0.0717 .
height        54.888      2.674  20.526   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.12 on 598 degrees of freedom
Multiple R-squared:  0.4133,    Adjusted R-squared:  0.4124
F-statistic: 421.3 on 1 and 598 DF,  p-value: < 2.2e-16

>
> abline(relation, cex = 1.3, pch = 16, xlab = "Height in meters", ylab = "Weight in kilos")
>
> prediction <- data.frame(height = 2.1)
> predict(relation, prediction)
       1
108.3021
```



Корелация съществува, но само докато височината стигне 2 метра. При покемон с височина от 2.1 метра, очакваме да тежи 108.3021 килограма.