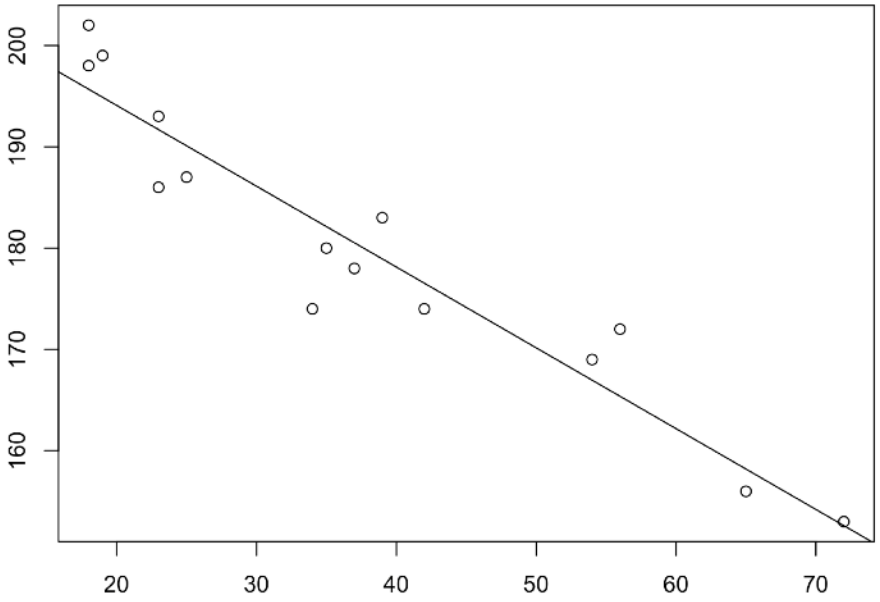
1. Начертана е графиката на данните "heartrade" от пакета "UsingR". Коя от възможностите може да отговаря на уравнението на регресия (коефициент на корелация, свободния член и коефициента пред x в уравнението на регресия)?

```
simple.lm(age, maxrate)
```




```
                    (Intercept)                   x
cor=-0.9534656, intercept=210.0485, heartrate$age=-0.7977
```

2. За данните "survey" от пакета "MASS" съдържащи височината на студентите "Height" - напишете функция, с която да направите проверка дали данните за височината на студентите са нормално разпределени. Получени са следните резултати. Анализирайте ги.

```
shapiro.test(Height)

        Shapiro-Wilk normality test

data:  Height
W = 0.98841, p-value = 0.08844
```

```
isNorm = function(X){
   sha = shapiro.test(X)
   p1 = sha$p.value

   if (p1 <= .05)
      print("not normal dist.")
   else
      print("normal dist.")
}
```

3. Интервалите между пристиганията на клиенти в работно време в магазин, измерени в часове са хи-квадрат разпределени със 3 степени на свобода.

   а) Симулирайте интервалите между пристиганията на 250 клиента, като вземете в предвид само работното време;

   б) Начертайте хистограма на данните;

   в) Сравнете хистограмата с теоритичното разпределение.

а) r=rchisq(n=250, df = 3)

б) hist(r, prob=T)

в)
```
points(min(r):max(r), dchisq(min(r):max(r),3), type="h", lwd=2, col="gold")
points(min(r):max(r), dchisq(min(r):max(r),3), type="p", lwd=3, col="purple")
curve(dchisq(x,3), add=T, col="purple", lwd=2)
```

4. Нека X е броя на пиките, паднали се при случайно изтегляне на 10 карти с връщане от колода от 52 карти. Използвайте вградени функции в R, за да:

   а) Генерирате 300 наблюдения над X ;

   б) Пресметнете теоритичната вероятност да има най-много 5 пики включително ;

   в) Пресметнете теоритичната вероятност, броят на изтеглените пики да попадне в интервала (2,7) ;

а) X = rbinom(n=300, size=10, prob=1/4)

б) pbinom(q=5, size=10, prob=1/4)

в)

I-ви начин:
```
pbinom(q=6, size=10, prob=1/4)-pbinom(q=2, size=10, prob=1/4)
```
II-ри начин:
```
pbinom(q=2, size=10, prob=1/4, lower.tail=F)-pbinom(q=6, size=10, prob=1/4,
lower.tail=F)
```
III-ти начин:
```
1-(pbinom(q=2, size=10, prob=1/4)+pbinom(q=6, size=10, prob=1/4, lower.tail=F))
```
IV-ти начин:
```
sol=function(s=10, p=1/4){
        ans=0
        for(i in 3:6){
                ans=ans+dbinom(x=i, size=s, prob=p)
        }
        ans
}
```

5. Клиентите на даден интернет доставчик създават нови акаунти със средна интензивност 10 Пресметнете какъв ще бъде средно броя на акаунтите, които ще бъдат създадени през утрешния ден.

   а) Симулирайте 200 реализации на X ;

   б) Пресметнете вероятността да бъдат създадени точно 13 акаунта ;

   в) Намерете минималния брой акаунти, които ще бъдат създадени утре с вероятност поне 95%.

а) sim=rpois(n=200, lambda=10)

б) dpois(x=13, lambda=10)

в)

I-ви начин:
```
qpois(.05, lambda=10)
```
II-ри начин:
```
qpois(.95, lambda=10, lower.tail=F)
```

6. Премахнете последните 3 колони в дейтасет survey.

```
c=ncol(survey)
left=c-2; right=c
s=survey[-c(left:right)]
```

7. Първите 6 реда на дейтасета "statistics" изглеждат така

	gender	martialStatus	workingStatus	age
1	Male	Married	Working	45
2	Female	Unmarried	Working	22
3	Female	Married	Not working	36
4	Male	Unmarried	Working	32
5	Female	Married	Working	42
6	Male	Unmarried	Not Working	28

Коя/Кои от следните функции в R ще добавят нова колона съдържаща нормализираната възраст

a. statistics$normalized <- age/sd(age)
b. statistics$normalized <- age-mean(age)/sd(age)
c. **statistics$normalized <- scale(age)** **единствения верен отговор**
d. statistics$normalized <- age-mean(age)

Друг начин, по който може да добавим колона с нормаличираната възраст е: statistics$normalized <- **(**age-mean(age, na.rm=T)**)**/sd(age, na.rm=T), забележете разликата с b. - тук делим на sd (стандартното отклонение) цялата разлика между дадена възраст и средната възраст

8. Първите 6 реда на data frame-а "statistics" изглеждат така (показано е на предходната задача)

Напишете код в R, който да конвертира пола във факторна променлива.

Отговор: as.factor(gender)

9. Първите 6 реда на дейтасета "ToothGrowth" изглеждат така

	len	supp	dose
1	4.2	VC	0.5
2	11.5	VC	0.5
3	7.3	VC	0.5
4	5.8	VC	0.5
5	6.4	VC	0.5
6	10.0	VC	0.5

Където "len" е дължината на зъба, "supp" е метода - витамин С или портокалов сок и "dose" е колко грама им е даден 0.5, 1 или 2 mg. Напишете скрипт в R, с който да нарисувате как се изменя дължината на зъба при промяна на дозата и метода.

Отговор:

```
xyplot(len~supp | dose)
```

10. Кое/Кои от следните разпределения не са асиметрични:

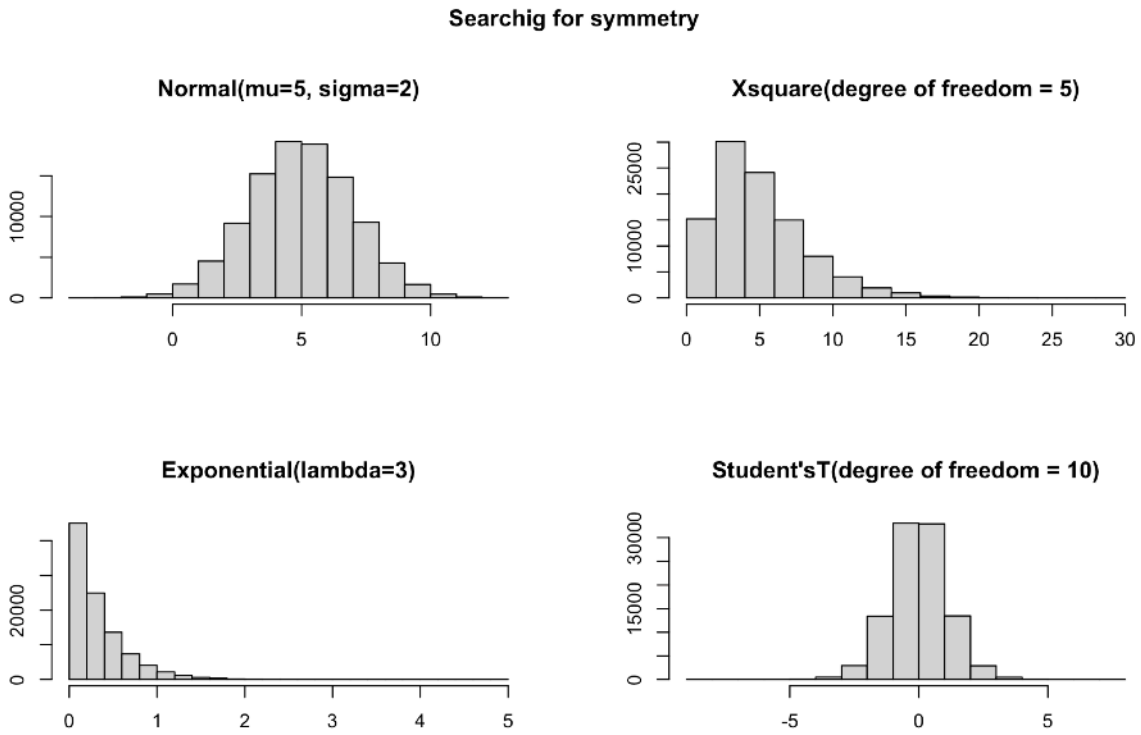
a. N(5, 2)
b. Chi^2(5)
c. Exp(3)
d. t(10)

```
sol=function(N=100000){
   X1 = rnorm(n=N, mean=5, sd=2)
   X2 = rchisq(n=N, df=5)
   X3 = rexp(n=N, rate=3)
   X4 = rt(n=N, df=10)

   par(mfrow=c(2,2),oma=c(0,0,2,0))

   hist(X1, main="Normal(mu=5, sigma=2)",xlab=NULL, ylab=NULL)
   hist(X2, main="Xsquare(degree of freedom = 5)",xlab=NULL, ylab=NULL)
   hist(X3, main="Exponential(lambda=3)",xlab=NULL, ylab=NULL)
   hist(X4, main="Student'sT(degree of freedom = 10)",xlab=NULL, ylab=NULL)

   title("Searchig for symmetry", outer=T)
}
```



Отговор: **a** и **d**

11. Първите 6 реда на дейтасета "statistics" изглеждат така

	gender	martialStatus	workingStatus	age
1	Male	Married	Working	45
2	Female	Unmarried	Working	22
3	Female	Married	Not working	36
4	Male	Unmarried	Working	32
5	Female	Married	Working	42
6	Male	Unmarried	Not Working	28

Какво ще изведе следния скрипт в R:

```
statistics[statistics$gender == "Male",][3, 4]
```

Отговор: Ще селектира само редовете на мъжете от дейтасет-а "statistics" и от тях ще изведе на колко години (4-та колона) е третия мъж (3-ти ред). Т.е. ще изведе числото **28.**

12. Напишете скрипт в R, който да конкатенира двете матрици по редове.

Отговор: `rbind(matrix1, matrix2)`

13. Първите 6 реда на дейтасета "students" изглеждат така

	Sex	Pulse	Exer	Smoke	Height	Age
1	Female		92	Some	Never	173.00	18.250
2	Male	104	None	Regul	177.80	17.583
3	Male	87	None	Occas	NA	16.917
4	Male	NA	None	Never	160.00	20.333
5	Male	35	Some	Never	165.00	23.667
6	Female		64	Some	Never	172.72	21.000

Напишете за всяка една от колоните качествени или количествени данни съдържа. Напишете скрипт в R, който да изведе честотната таблица на това колко често пушат студентите и направете подходяща графика.

Отговор:

Sex - качесвени
Pulse - дискретни количествени (числови)
Exer - качествени
Smoke -качествени
Height - непрекъснати количествени (числови)
Age - непрекъснати количествени (числови)

```
prob.table(table(Smoke))
barplot(prop.table(table(Smoke)))
```

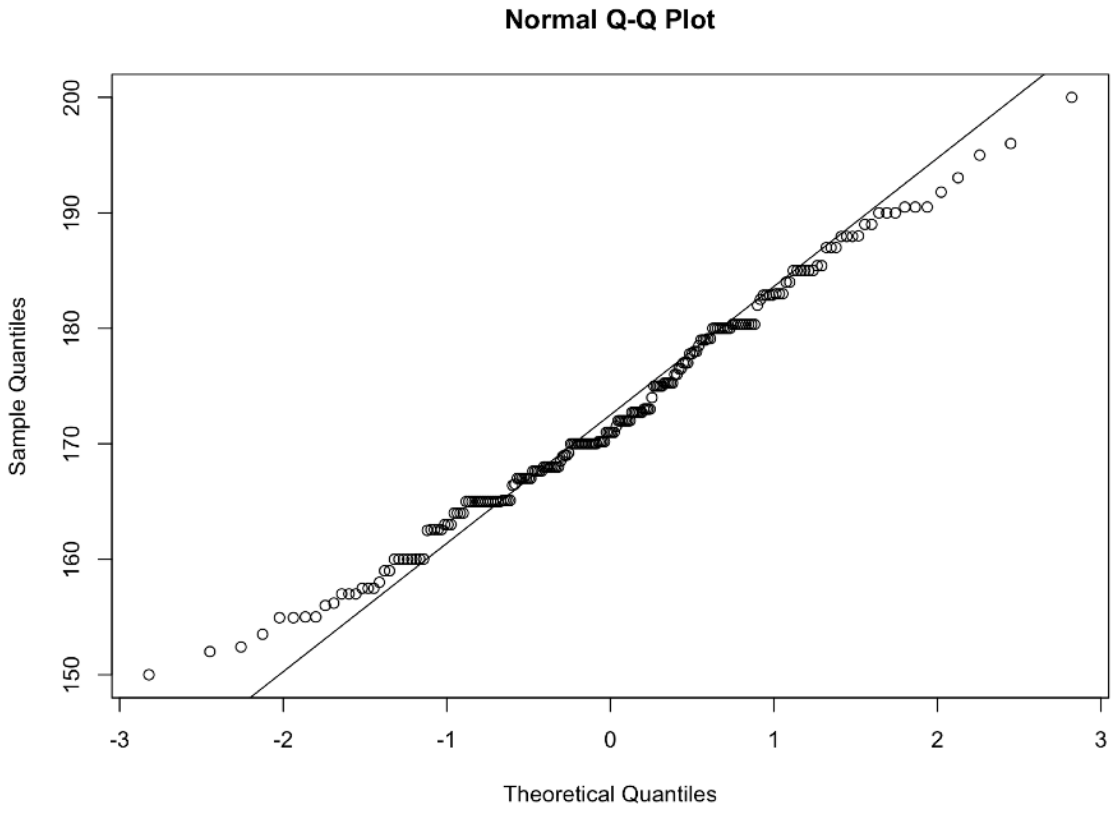
14. Мария за домашното си по география трябвало да намери данни за няколко страни и географски показатели за тях. Тя ги записвала в дейтасет "countries" и първите 6 реда от таблицата ѝ изглеждали по следния начин:

	country	population2020	landarea	density
1	Austria	9006398	82409	109
2	Bhutan	NA	NA	20
3	Canada	37742154	9093510	4
4	Ethiopia	NA	1000000	NA
5	Finland	5540720	NA	NA
6	Honduras	NA	111890	89

Напишете скрипт в R, с който да и помогнете да преброи общия брой липсващи стойности "NA" , които ѝ остава да попълни.

Отговор: sum(is.na(countries))

15. За данните от "survey" от пакета "MASS" съдържащи височините на стундентите "Height" - напишете функция, с която да направите qq-plot за височината на студентите. Получена е следната графика. Анализирайте графиката.



R код:
```
library(MASS)
attach(survey)
qqnorm(Height)
qqline(Height)
```

qqplot-а сравнява квантила на данните ни с квантила на разпределението което му зададем като второ. В случая използваме вградената функция qqnorm, която знае че ще сравнява данните ни с нормалното разпределение. Когато имаме разлики само в средното - линията ще се транслира нагоре или надолу. Когато имаме разлики само в дисперсията - линията ще се завърти в центъра и в опашките ще се получи раздалечаване от нея. Точно това се получава и на нашата графика - т.е. височината на студентите е нормално

разпределена, с разликата че има по голямо стандартно отклонение отколкото би имало нормалното разпределение със средно, което е равно на средното на нашите данни и дисперсия, която е равна на дисперсията на нашите данни. С други думи имаме по-голяма вероятност да срещнем по-далечно наблюдение в нашите данни, отколкото в данните на нормалното разпределение. В случай, че имаме разлики и в средното и в дисперсията - линията изобщо няма да съответства по никакъв начин на данните ни. Ако например сравним нашите данни за височината с експоненциалното разпределение ще видим, че линията и данните няма да имат нищо общо.

Ако искаме да сравним с някакво друго разпределение, а не с нормалното и отново да е поквантилно, то тогава може да използваме qqplot.das(X, "norm"), като например заместим нормалното с експоненциално и т.н. qqplot.das(X, "exp").

16. Първите 6 реда на "Students" изглеждат както в зад. 13.

Напишете скрипт в R, който въз основа на данните да оцени вероятността случайно избран, понякога спортуващ човек, да се окаже с пулс между 60 и 70 удара в минута включително. Начертайте хистограми на разпределението на студентите според пулса им, ако е известно колко често спортуват.

```
# students=survey

some = survey[!is.na(survey$Exer) & survey$Exer == "Some",]
pulse = survey[!is.na(survey$Pulse) & survey$Pulse >= 60 & survey$Pulse <= 70,]
P=(nrow(pulse) / nrow(some)) * 100; P

# hist
par(mfrow=c(3,1))
hist(survey[survey$Exer == "Some",]$Pulse, prob = T)
hist(survey[survey$Exer == "None",]$Pulse, prob = T)
hist(survey[survey$Exer == "Freq",]$Pulse, prob = T)
```