

## 1. R eta RStudio: ezaugarri nagusiak

### 1. ARIKETA

**dataset** paketearen barnean dagoen **quakes** datu-markoa kontsideratuko da. Bertan, Fiji inguruan izandako lurrikarak aztertzen dira.

- Instalatu eta kargatu **dataset** paketea eta aztertu **quakes** datu-markoak dituen behaketa eta aldagai kopuruak.
- Aztertutako lurrikaretatik zenbatetan gorputz uhinen magnitudea 6 mB baino handiagoa da?
- Lortu gutxienez 6 mB-ko gorputz uhinen magnitudea eta -16 latitudetik iparralderago gertatutako lurrikaren informazioa.
- Lortu gutxienez 5 mB-ko gorputz uhinen magnitudea eta bere fokua 600 km baino sakonago dituzten lurrikaren latitudea, longituda, eta informazioa jaso duen estazioaren zenbakiaren identifikatzailea.

---

Ebazpena :

a)

**datasets** paketea instalatu eta kargatuko dugu:

```
> install.packages("datasets")  
> library(datasets)
```

str() funtzioak datu-markoaren egitura bueltatzen du. **quakes** datu-markoak 5 aldagai eta 1000 behaketa ditu:

```
> str(quakes)  
'data.frame': 1000 obs. of 5 variables:  
 $ lat      : num  -20.4 -20.6 -26 -18 -20.4 ...  
 $ long     : num   182 181 184 182 182 ...  
 $ depth    : int   562 650 42 626 649 195 82 194 211 622 ...  
 $ mag      : num   4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...  
 $ stations: int    41 15 43 19 11 12 43 15 35 19 ...
```

b)

```
> quakes$mag[quakes$mag>6]  
[1] 6.1 6.4
```

```
> Length(quakes$mag[quakes$mag>6])
[1] 2
```

Beraz, aztertutako 1000 lurrikaretatik bakarrik bitan gorputz uhinen magnitudea 6 mB baino handiagoa da.

c)

```
> quakes[quakes$mag>=6 & quakes$lat > -16, ]
      lat  long depth mag stations
17  -13.64 165.96   50 6.0       83
152 -15.56 167.62  127 6.4      122
870 -12.23 167.02  242 6.0      13
```

d)

Datu-markoaren egitura ikusi bezala, latitudea, longituda, eta estazioaren zenbakiaren identifikatzailea lehenengo, bigarren eta bosgarren aldagaiak dira:

```
> quakes[quakes$mag>=5 & quakes$depth > 600, c(1,2,5)]
      lat  long stations
63  -13.82 172.38      61
68  -21.96 179.62      45
207 -21.96 180.54      66
238 -21.68 180.63      63
291 -18.48 181.49      49
312 -12.93 169.63      57
374 -19.77 181.40      54
385 -20.04 182.01      49
400 -13.45 170.30      93
462 -18.00 180.62     100
463 -18.08 180.70      72
651 -18.12 181.88      88
663 -18.14 180.87     105
666 -18.21 180.87      69
697 -18.07 181.58      65
724 -26.50 178.29      50
753 -21.08 180.85     119
```

## 2. ARIKETA

**datasets** paketearen barnean dagoen **rivers** bektorea kontsideratu:

- Instalatu eta kargatu **datasets** paketea eta azaldu **rivers** bektorea.
- histograma batean **rivers** bektorearen balioak irudikatu.
- rivers** bektorearen kutxa diagrama irudikatu eta balio arraroak lortu.
- rivers** bektorean agertzen diren erreken batezbesteko luzera kalkulatu balio arraroak kendu gabe eta balio arraroak kenduta.

---

Ebazpena :

a)

**datasets** paketea instalatu eta kargatuko dugu:

```
> install.packages(rivers)
> library(datasets)
```

**rivers** bektorea azaltzeko laguntza erabiliko dugu (`help()` funtzioa edo `?eragilea`):

```
> ?rivers # edo help(rivers)
```

rivers {datasets}	R Documentation
<h3>Lengths of Major North American Rivers</h3>	
<b>Description</b>	
This data set gives the lengths (in miles) of 141 "major" rivers in North America, as compiled by the US Geological Survey.	
<b>Usage</b>	
<code>rivers</code>	
<b>Format</b>	
A vector containing 141 observations.	
<b>Source</b>	
World Almanac and Book of Facts, 1975, page 406.	
<b>References</b>	
McNeil, D. R. (1977) <i>Interactive Data Analysis</i> . New York: Wiley.	

**rivers** bektoreak Ipar Amerikako 141 erreka nagusien luzera (militan) gordetzen ditu.

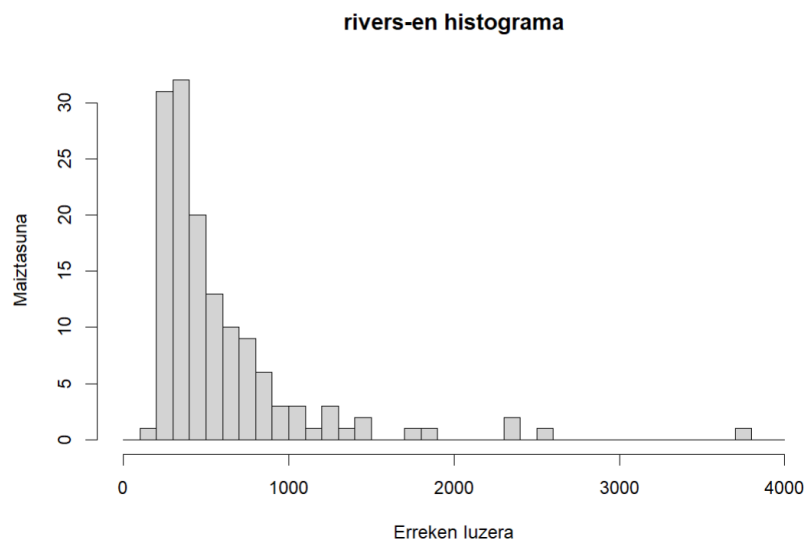
b)

Lehenik, aldagaiaren heina lortuko dugu:

```
> range(rivers)
[1] 135 3710
```

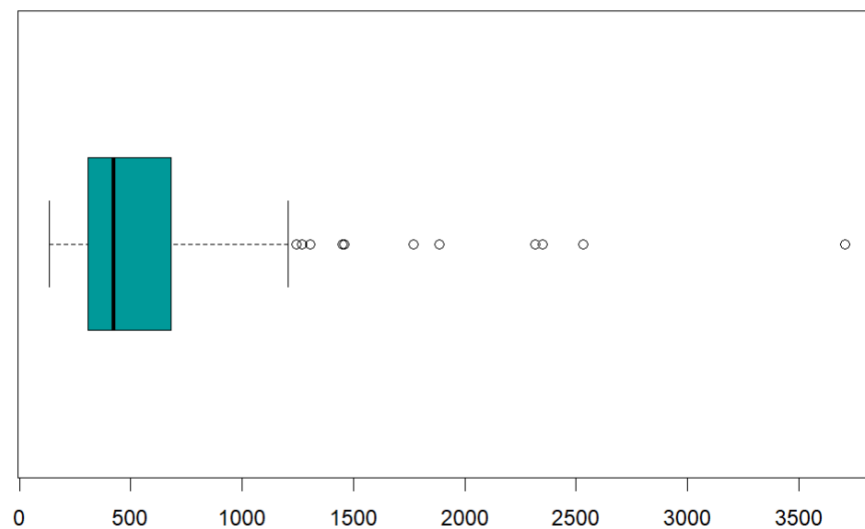
Histograma egiteko 0-tik 4000-rako 100 tarte berdinean erabiliko ditugu:

```
> hist(rivers,breaks=seq(0,4000,100),xlab="Erreken  
luzera",ylab="Maiztasuna", main="rivers-en histograma")
```



c)

```
> boxplot(rivers, horizontal = T, col="#009999")
```



```

> boxplot(rivers, horizontal = T, col="#009999"
$stats
[1] 135  310  425  680 1205

$n
[1] 141

$conf
[1] 375.7678 474.2322

$out
[1] 1459 1450 1243 2348 3710 2315 2533 1306 1270 1885
[11] 1770
  
```

**rivers** bektorearen balio arrakoak *out* aldagaian agertzen dira: 1459, 1450, 1243, 2348, 3710, 2315, 2533, 1306, 1270, 1885 eta 1770 milia dira.

d)

Bektoreen balioen batezbestekoa lortzeko `mean()` funtzioa erabiliko dugu:

```

> mean(rivers)
[1] 591.1844
  
```

```

> boxplot.stats(rivers)$out
[1] 1459 1450 1243 2348 3710 2315 2533 1306 1270 1885 1770
> which(rivers %in% boxplot.stats(rivers)$out)
[1]  7  23  25  66  68  69  70  83  98 101 141
> x[-which(rivers %in% boxplot.stats(rivers)$out)]
[1] 17 23 21 36 21 18 22 35 20
> rivers.na.gabe<-
rivers[-which(rivers %in% boxplot.stats(rivers)$out)]
[1]  735  320  325  392  524  450  135  465  600
[10]  330  336  280  315  870  906  202  329  290
[19] 1000  600  505  840  890  350  407  286  280
[28]  525  720  390  250  327  230  265  850  210
[37]  630  260  230  360  730  600  306  390  420
[46]  291  710  340  217  281  352  259  250  470
  
```

```
[55] 680 570 350 300 560 900 625 332 1171
[64] 780 280 410 460 260 255 431 350 760
[73] 618 338 981 500 696 605 250 411 1054
[82] 735 233 435 490 310 460 383 375 545
[91] 445 380 300 380 377 425 276 210 800
[100] 420 350 360 538 1100 1205 314 237 610
[109] 360 540 1038 424 310 300 444 301 268
[118] 620 215 652 900 525 246 360 529 500
[127] 720 270 430 671
> mean(rivers.na.gabe)
[1] 477.4462
```

Balio arraroak kendu ondoren batezbestekoa txikitu egiten da. Argi geratzen da muturreko balioek batezbestekoan daukaten eragina.

### 3. ARIKETA

**carData** paketearen barnean dagoen Florida data-markoa kontsideratuko da. Data-marko honek 2000. urteko hauteskundeetan Floridako konderrietan hautagai bakoitzak izandako boto kopurua erakusten du. Honakoa eskatzen da:

- carData** paketea instalatu eta kargatu
- Floridako zenbat konderritan Al Gore gailendu zitzaion George Bush-i?
- Izendatu Al Gore George Bush-i gailendutako konderriak.
- Zenbat boto jaso zuen George Bush-ek guztira Floridako estatuan? Eta Al Gore-k?
- Zenbat boto jaso zuen hautagai bakoitzak BAY eta CHARLOTTE konderrietan?
- Zein konderritan jaso zituen George Bush-ek boto gehien? Eta Al Gore-k?
- Irudikatu boto gehien lortutako bost hautagaien barra diagrama.

Ebazpena :

a)

```
> install.packages(carData)
> library(carData)
```

b)

```
> which(Florida$"GORE">Florida$"BUSH")
[1] 1 6 13 18 20 27 33 37 44 48 49 50 51 52 56 64
> length(which(Florida$"GORE">Florida$"BUSH"))
[1] 16
```

Al Gore Floridako 16 konderritan gailendu zitzaion George Bush-i.

c)

Konderrien izenak lortzeko rownames() funtzioa erabiliko da:

```
> rownames(Florida)[which(Florida$"GORE">Florida$"BUSH")]
[1] "ALACHUA" "BROWARD" "DADE" "FLAGLER" "GADSDEN"
[6] "HERNANDO" "JEFFERSON" "LEON" "MONROE" "ORANGE"
[11] "OSCEOLA" "PALM.BEACH" "PASCO" "PINELLAS" "ST.LUCIE"
[16] "VOLUSIA"
```

d)

```
> sum(Florida$"BUSH")
[1] 2904836
> sum(Florida$"GORE")
[1] 2903861
```

George Bush-ek 2.904.836 boto jaso zituen Floridako estatuan. Al Gore-k, aldiz, 2.903.861 boto.

e)

“BAY” eta “CHARLOTTE” errenkadetako aldagai guztiak bistaratuko ditugu:

```
> Florida[c("BAY", "CHARLOTTE"), ]
```

	GORE	BUSH	BUCHANAN	NADER	BROWNE	HAGELIN
BAY	18850	38637	248	828	171	18
CHARLOTTE	29641	35419	182	1461	127	15
	HARRIS	MCREYNOLDS	MOOREHEAD	PHILLIPS	Total	
BAY	5	3	37	18	58815	
CHARLOTTE	6	3	12	19	66885	

f)

```
> which(Florida$BUSH==max(Florida$BUSH))
[1] 13
> Florida$BUSH[13]
[1] 289456
> rownames(Florida)[13]
[1] "DADE"
```

```
> which(Florida$GORE==max(Florida$GORE))
[1] 6
> Florida$GORE[6]
[1] 386518
> rownames(Florida)[6]
[1] "BROWARD"
```

George Bush-ek DADE konderrian lortu zituen boto gehien. Al Gorek, aldiz, BROWARD konderrian.

g)

Lehenengo, hautagai guztiek lortutako boto kopurua lortuko dugu:

```
> apply(Florida, 2, sum)
```

GORE	BUSH	BUCHANAN	NADER	BROWNE	HAGELIN
2903861	2904836	17317	96563	18808	2285
HARRIS	MCREYNOLDS	MOOREHEAD	PHILLIPS	Total	
10471	1275	1839	4276	5961531	



Hautagaiak boto kopuruaren arabera ordenatuko ditugu, ordena beherakorrean:

```
> sort(apply(Florida,2,sum),decreasing=T)
  Total      BUSH      GORE      NADER      BROWNE      BUCHANAN
5961531  2904836  2903861    96563    18808    17317

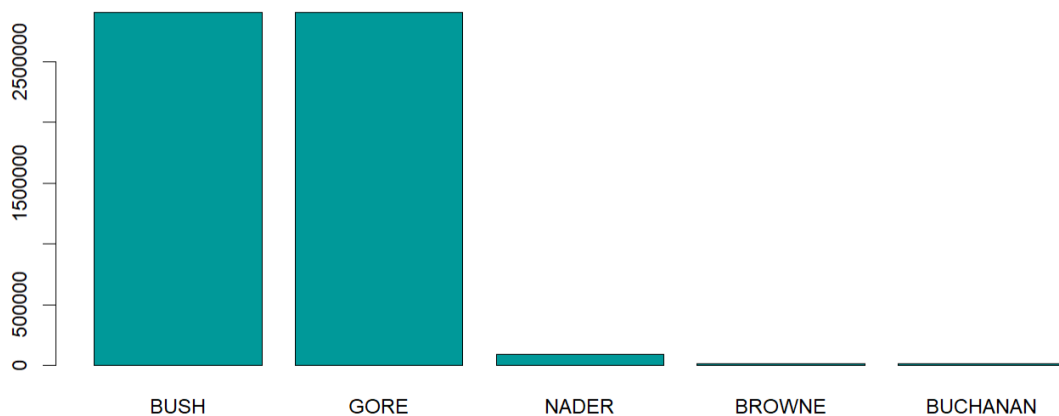
  HARRIS  PHILLIPS  HAGELIN  MOOREHEAD  MCREYNOLDS
  10471    4276      2285      1839      1275
```

Lehenengo aldagaia boto kopuru totala da. Horrela, boto gehien lortutako bost hautagaien botoak lortzeko:

```
> sort(apply(Florida,2,sum),decreasing=T)[2:6]
      BUSH      GORE      NADER      BROWNE  BUCHANAN
2904836  2903861    96563    18808    17317
```

Azkenik, barra diagrama irudikatzeko barplot() funtzioa erabiliko da:

```
> barplot(sort(apply(Florida,2,sum),decreasing=T)[2:6],
  col="#009999")
```



#### 4. ARIKETA

**datasets** paketearen barnean dagoen **airquality** datu-markoa kontsideratuko da. Datu-marko honek 1973ko maiatza eta iraila hilabeteen artean New York hirian egunero neurtutako ozono maila, eguzki-erradiazio maila, haize-abiadura, eta temperatura maximoa erakusten ditu. Honakoa eskatzen da:

- Instalatu eta kargatu **datasets** paketea
- Lortu **airquality** datu-markoaren aldagaien laburpena (minimoa, 1. kuartila, mediana, batezbestekoa, 3. kuartila eta maximoa).
- Bistaratu datu-markoaren lehenengo zortzi behaketen emaitzak.
- Lortu eguzki-erradiazio mailaren batezbestekoa. Ezabatu balio galduren bat (NA-not available) duten behaketak eta kalkulatu berriro eguzki-erradiazio mailaren batezbestekoa.
- Laguntza erabiliz aurkitu zein unitatetan adierazi diren aldagaien balioak.
- Aldatu aldagaien unitateak: ozono maila ppm-tan, eguzki-erradiazio maila  $\text{kJ/m}^2$ -tan, haize-abiadura km/ordu-tan eta temperatura gradu Celsius-tan. Lortutako aldagaiekin sortu **airquality2** datu-markoa eta bistaratu lehenengo 8 behaketak.

---

Ebazpena :

a)

datasets paketea instalatu eta kargatuko dugu:

```
> install.packages(rivers)
> library(datasets)
```

b)

Laburpena lortzeko `summary()` funtzioa erabiliko dugu:

```
> summary(airquality)
```

Ozone	Solar.R	wind
Min. : 1.00	Min. : 7.0	Min. : 1.700
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400
Median : 31.50	Median :205.0	Median : 9.700
Mean : 42.13	Mean :185.9	Mean : 9.958
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500
Max. :168.00	Max. :334.0	Max. :20.700
NA's :37	NA's :7	

Temp	Month	Day
Min. :56.00	Min. :5.000	Min. : 1.0
1st Qu.:72.00	1st Qu.:6.000	1st Qu.: 8.0
Median :79.00	Median :7.000	Median :16.0
Mean :77.88	Mean :6.993	Mean :15.8
3rd Qu.:85.00	3rd Qu.:8.000	3rd Qu.:23.0
Max. :97.00	Max. :9.000	Max. :31.0

c)

Lehenengo zortzi behaketak bistartzeko head() funtzioa erabiliko dugu:

```
> head(airquality,8)
```

	Ozone	Solar.R	wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	5

d)

summary() funtzioa erabiliz lortutako laburpenean balio galduak daudela ikus daiteke. Ozone aldagaiak 37 balio galdu ditu eta Solar.R aldagaiak 7 balio galdu ditu. Honela, mean() funtzioa erabili ahal izateko aldagaiak dituen balio galduak ezabatzea beharrezkoa da:

```
> mean(airquality$Solar.R)
[1] NA
> mean(airquality$Solar.R, na.rm=T)
[1] 185.9315
```

Datu-markotik balio galduren bat dituzten behaketak ezabatzeko na.omit() funtzioa erabiliko dugu. Ondoren mean() funtzioarekin Solar.R aldagaiaren batezbesteko balioa lortuko dugu:

```
> airquality.na.gabe<-na.omit(airquality)
> mean(airquality.na.gabe$Solar.R)
[1] 184.8018
```

e)

Aldagaiari buruzko laguntza lortzeko help() funtzioa edo ? operadorea erabili daitezke

```
> help(airquality) # edo ?airquality
```

airquality {datasets}

R Documentation

## New York Air Quality Measurements

### Description

Daily air quality measurements in New York, May to September 1973.

### Usage

```
airquality
```

### Format

A data frame with 153 observations on 6 variables.

```
[,1] Ozone    numeric Ozone (ppb)  
[,2] Solar.R  numeric Solar R (lang)  
[,3] Wind     numeric Wind (mph)  
[,4] Temp    numeric Temperature (degrees F)  
[,5] Month   numeric Month (1--12)  
[,6] Day     numeric Day of month (1--31)
```

Ozono maila mila milioiko partetan (ppb), eguzki erradiazioa lang-etan, haizearen abiadura orduko miliatan (mph) eta tenperatura Fahrenheit gradutan adierazita daude.

f)

```
> attach(airquality)  
> Ozone.ppm<-Ozone/1000  
> Solar.R.kJ.m2<-Solar.R*2.4*10^(-2)  
> Wind.km.o<-wind*1.60934  
> Temp.C<-(Temp-32)/1.8  
> airquality2<-data.frame(Ozone.ppm, Solar.R.kJ.m2,  
Wind.km.o, Temp.C)
```

Lehenengo 8 behaketak bistartzeko berriro ere head() funtzioa erabiliko dugu:

```
> head(airquality2,8)
  Ozone.ppm solar.R.kJ.m2 wind.km.o Temp.C
1    0.041      4.560  11.90912  19.44444
2    0.036      2.832  12.87472  22.22222
3    0.012      3.576  20.27768  23.33333
4    0.018      7.512  18.50741  16.66667
5         NA         NA  23.01356  13.33333
6    0.028         NA  23.97917  18.88889
7    0.023      7.176  13.84032  18.33333
8    0.019      2.376  22.20889  15.00000
```

## 5. ARIKETA

**palmerpenguins** paketearen barnean dagoen **penguins** datu-markoa kontsideratuko da. Palmer uhartediko hiru uhartetan (Biscoe, Dream eta Torgersen) aurkitutako hiru pinguino espezieri (Adelie, Chinstrap eta Gentoo) buruzko gorputzaren atalak (mokoaren luzera eta sakonera, hegalen luzera eta pisua) aztertu dira:

- Instalatu eta kargatu **palmerpenguins** paketea
- Bistaratu **penguins** datu-markoa.
- Sektore diagrama irudikatu pinguinoen espezie aldagaia kontutan hartuz.
- Irudikatu barra diagrama bat uharte bakoitzean dauden pinguino espezieak bistaratzeko.
- Lortu 5000 g baino pisu gehiago dituzten pinguinoen artetik hegal laburrenaren luzera.
- Zein uhartetan aurkitu da moko luzeena duen pinguinoa?

Ebazpena :

a)

palmerpenguins paketea instalatu eta kargatuko dugu:

```
> install.packages(palmerpenguins)
> library(palmerpenguins)
```

b)

Datu-markoa bistaratzeko View() funtzioa erabiliko dugu:

> View(penguins)

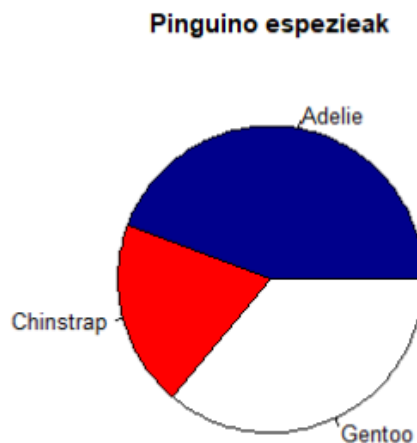
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

Showing 1 to 11 of 344 entries, 8 total columns

c)

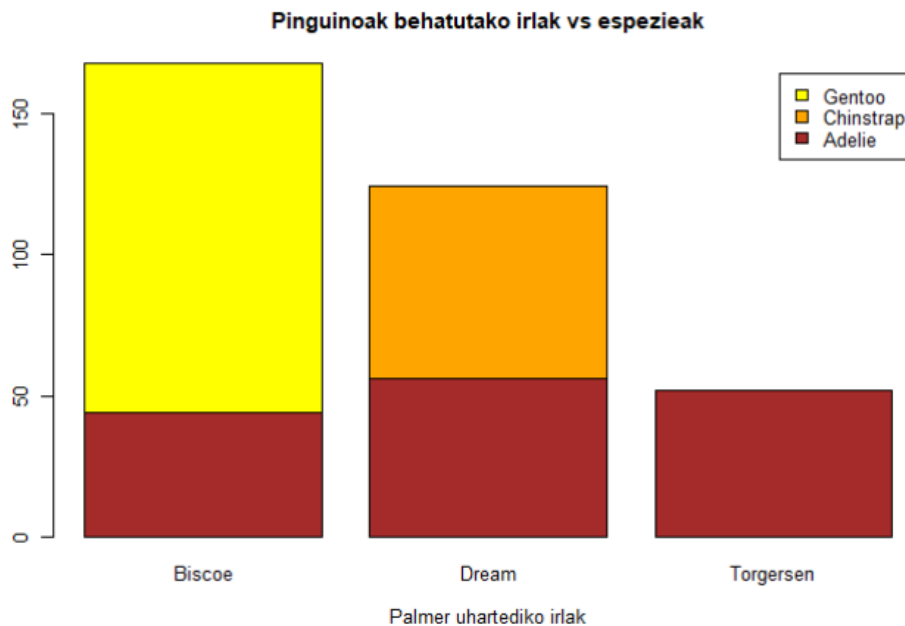
sektore-diagrama irudikatzeko pie() funtzioa erabiliko dugu:

```
> pie(table(penguins$species),col=c("darkblue","red",
"white"),main="Pinguino espezieak")
```



d)

```
> barplot(table(penguins$species,penguins$island),
main="Pinguinoak behatutako irlak vs espezieak",
xlab="Palmer uhartediko irlak",
col=c("brown","orange","yellow"),
legend=rownames(table(penguins$species)))
```



Barra diagraman ikusten denez, Adelie espezieko pinguinoak hiru uharteetan aurkitu dira. Aldiz, Gentoo espezie motako pinguinoak Biscoe uhartean aurkitu dira bakarrik. Gauza bera gertatu da Chinstrap espezie motarekin, Dream uhartean bakarrik aurkitu baita.

e)

Balio galduak dituzten behaketak kenduko ditugu lehenik eta behin. Bestela hegala laburrena NA (balio galdua) izango zen.

```
> penguins.na.gabe<-na.omit(penguins)
> min(penguins.na.gabe$flipper_length_mm
[penguins.na.gabe$body_mass_g>=5000])
[1] 207
```

5000 g baino pisu gehiago dituzten pinguinoen artean hegala laburrenaren luzera 207 mm da.



f)

```
> subset(penguins.na.gabe,  
penguins.na.gabe$bill_length_mm==59.6)  
# A tibble: 1 x 8  
  species island bill_length_mm bill_depth_mm  
flipper_length_mm body_mass_g sex year  
186 Gentoo Biscoe          59.6           17  
                230          6050 male 2007
```

Moko luzeena duen pingüinoa Biscoe uhartean aurkitu zen.