

R softwarea: mapen bisualizazioa eta parametro estatistikoen kalkulua

Gorka Kobeaga ^{a,b}

^aBilbaoDataLab

^bBCAM - Basque Center for Applied Mathematics

2017.eko azaroaren 17
#Geomatika & Inteligentzia espaziala



EDUKIAK

Sarrera

Mapak

Datuak

Klusterrak

ZER DA R?

Analisi estadistiko eta grafikorako, programazio hizkuntza eta ingurune bat da R. Erabilera desberdinetarako pakete eta liburutegiak kargatu daitezke.

- ▶ R 1993an hasi zen garatzen Auckland-eko Unibertsitatean (Zelanda Berria).
- ▶ Software Libreko proiektu bat da. GNU/GPL lizentziapean banatzen da.
- ▶ Bere aitzindaria S izan zen. Objetuetara bideratutako programazio hizkuntzaren ezaugarriak ditu.
- ▶ Python eta Perl bezalako hizkuntza interpretatuetan erabili daiteke.
- ▶ GNU/Linux, Windows, Macintosh, Unix-erako dago.

FITXATEGIAK IRAKURTZEA

▶ Taulak

```
> read.table(file="fitxategia", header = FALSE,  
  sep = "", dec = ".", na.strings = "NA")
```

▶ CSV

```
> read.csv(file="fitxategia", header = TRUE, sep = ",",  
  dec=".")
```

▶ SAV

```
> library(foreign)  
> read.spss(file="fitxategia", use.value.labels = TRUE)
```

BEKTOREAK

▶ Bektoreak sortzeko:

```
> bektorea1 <- c(1,2,3,4)
> bektorea2 <- c(7:10) # c(7,8,9,10)-ren baliokidea
> bektorea3 <- rep(5,2) # c(5,5)-ren baliokidea
> bektorea4 <- seq(1,10,2) # c(1, 3, 5, 7, 9)-ren baliokidea
```

▶ Baloioak gehitzeko edo bektoreak bitzeko:

```
> bektorea3 <- c(bektorea2,bektorea1)
> bektorea3
```

▶ Balioak aukeratzeko:

```
> bektorea3[c(4:6)] # 4., 5. eta 6. balioak aukeratu
```

MATRIZEAK

```
> matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,  
dimnames = NULL)
```

```
> matizeal <- matrix(c(1,2,3,4,5,6),nrow=2,ncol=3)  
> matizeal
```

	[1,]	[2,]	[3,]
[1,]	1	3	5
[2,]	2	4	6

MATRIZEAK

- ▶ Zutabeak gehitzeko:

```
> matrizea2 <- cbind(matrizea1,c(8,NA)); matrizea2
```

	[1]	[2]	[3]	[4]
[, 1]	1	3	5	8
[, 2]	2	4	6	NA

- ▶ Errenkadak gehitzeko:

```
> matrizea3 <- rbind(matrizea1,c(8,10,-1)); matrizea3
```

- ▶ Matrize iraulia:

```
> t(matrizea3)
```

	[1]	[2]	[3]
[, 1]	1	2	8
[, 2]	3	4	10
[, 3]	5	6	-1

BESTE KOMANDO ERABILGARRI BATZUK

Liburutegiak

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

Laguntza

```
> ?plot
```

NA: Balio galduak

- ▶ Jakiteko zenbat balio galdu dauden bektore batean:

```
> sum(is.na(matrizea2[,4]))  
[1] 1
```

- ▶ Jakiteko non dauden balio galduak bektore batean:

```
> which(is.na(matrizea2[,4]))  
[1] 2
```


BESTE KOMANDO ERABILGARRI BATZUK

Ordenatzeko

- ▶ **sort** Bektore bat ordenatzeko sort komando erabili daiteke:

```
> sort(c(2:3, 0, 4,1))  
[1] 0 1 2 3 4
```

- ▶ **order** Balioen ordena itzultzen du bektore batean:

```
> order(c(2:3, 0, 4,1))  
[1] 3 5 1 2 4
```

- ▶ **order** Matrize bateko lerroak ordenatzeko erabili daiteke:

```
> matricea3[order(matricea3[,3]),]  
      [,1] [,2] [,3]  
[ ,1]  8   10  -1  
[ ,2]  1    3   5  
[ ,3]  2    4   6
```

DATU MULTZOAK

R-k berez hainbat datu multzo ditu sartuta.

```
> data() # datu multzoak ikusteko
```

Datu multzo bat kargatzeko eta arakatzeko:

```
> data(mtcars)
> ?mtcars # Laguntzan: aldagaiei buruzko informazioa
> str(mtcars) # Datu multzoaren estruktura
'data.frame': 32 obs. of 11 variables:
 $ mpg   : num   21  21  22.8  21.4  18.7  18.1 ...
 $ cyl   : num    6  6  4  6  8  6 ...
 $ disp  : num  160 160 108 258 360 ...
```

```
> dim(rnd.matrizea) # nrow eta ncol
```

```
[1] 32 11
```

```
> head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875
Datsun 710	22.8	4	108	93	3.85	2.320

SUMMARY

```
> summary(mtcars)
```

	mpg	cyl	disp
Min. :	10.4	4.00	71.1
1st Qu.:	15.4	4.00	120.8
Median :	19.2	6.00	196.3
Mean :	20.1	6.19 6	230.7
3rd Qu.:	22.8	8.00	326.0
Max. :	33.9	8.00	472.0

GRAFIKOAK

Grafikoak egiteko R-ren oinarrizko paketea **graphics** da: **plot**, **contour** eta **perps** bezalako funtzioak bertan bilduta daude. Ereduak ikusteko:

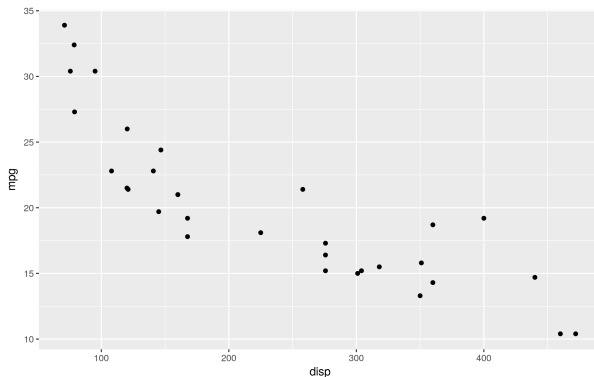
```
> demo(graphics)
> demo(image)
> demo(perps)
```

Hala ere, tailer honetan, grafikoak egiteko **ggplot2**¹ paketea erabiliko dugu:

```
> library(ggplot2)
```

¹<https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

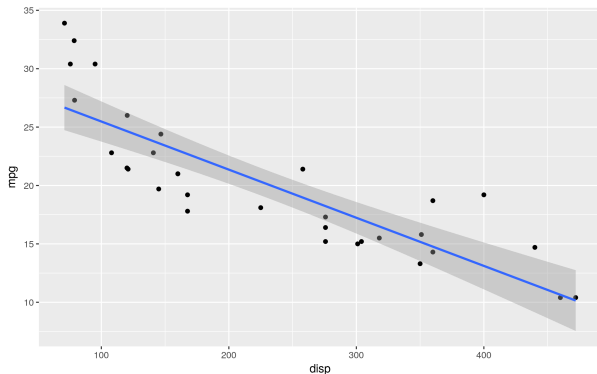
GRAFIKOAK



```
> ggplot(data=mtcars  
  , aes(x=disp,y=mpg))  
  geom_point()
```

Datuak
+ # Itxura
Puntuak

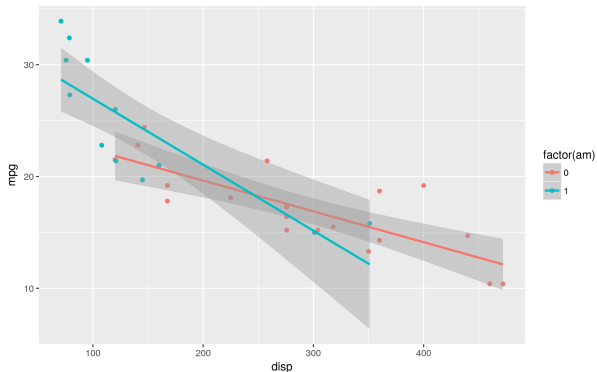
GRAFIKOAK



```
> ggplot(data=mtcars  
          , aes(x=disp,y=mpg))  
  geom_point()  
  stat_smooth(method="lm")
```

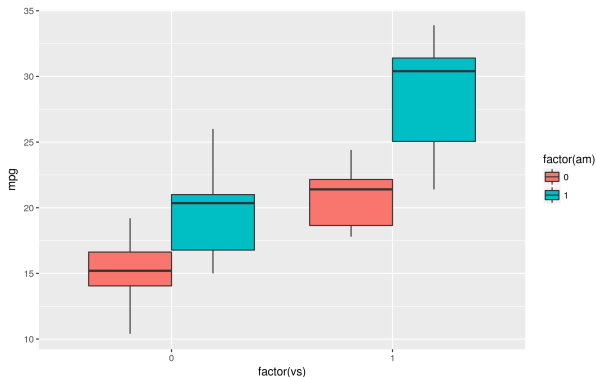
Datuak
+ # Itxura
+ # Puntuak
Erregresio lineala

GRAFIKOAK



```
> ggplot(data=mtcars, # Datuak  
  aes(x=disp,y=mpg,color=factor(am))) + # Itxura  
  geom_point() + # Puntuak  
  stat_smooth(method="lm") # Erregresio lineala
```

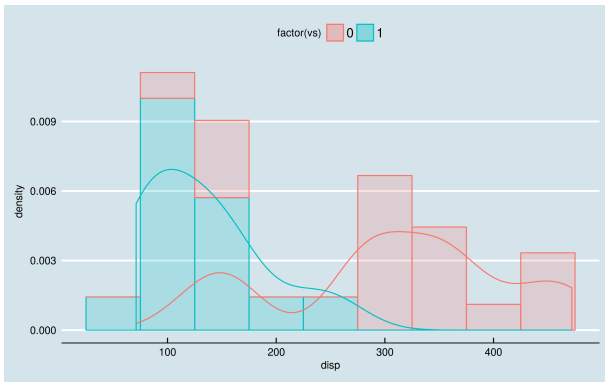
GRAFIKOAK



```
> ggplot(data=mtcars,  
  aes(x=factor(vs), y=mpg, fill=factor(am))) +  
  stat_boxplot()
```

Datuak
Itxura
Kutxak

GRAFIKOAK

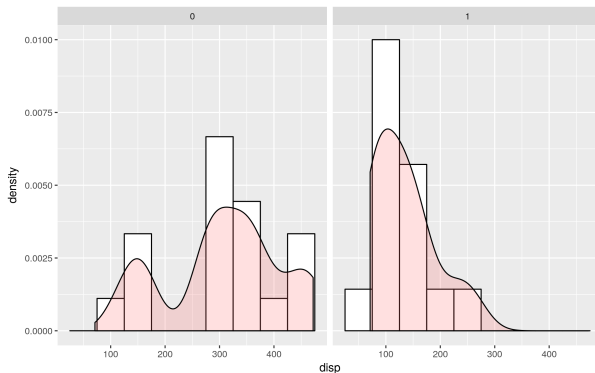


```

> g <- g + ggplot(mtcars, aes(x=disp), color=factor(vs))
> g <- g + geom_histogram(aes(y=..density..) # Histograma
, fill=factor(vs), binwidth=50, alpha=0.3)
> g <- g + geom_density(alpha=.2) + # Dentsitatea
> g <- g + theme_economist()
> gg

```

GRAFIKOAK



```

> g <- g + ggplot(mtcars, aes(x=disp))
> g <- g + geom_histogram(aes(y=..density..),
  binwidth=50, colour="black", fill="white")
> g <- g + geom_density(alpha=.2,fill="#FF6666")
> g <- g + facet_wrap(~ vs) # Grafiko anizkoitza sortzeko
> g

```

LAGUNTZA

Hurrengo orrialdeetan arazo eta galdera askoren erantzuna aurkituko duzu:

- ▶ Rstudio-ren laburpenak
<https://www.rstudio.com/resources/cheatsheets/>
- ▶ Stack Overflow
<https://stackoverflow.com/questions/tagged/r>
- ▶ Nabble:
<http://r.789695.n4.nabble.com>

ARIKETAK I

1. Eman begirada bat R -n eskuragarri dauden datu-multzoei. Aukeratu bat.
2. Summary erabiliz kalkulatu aldagaien neurri deskribatzaileak.
3. Datu multzo horrentzat, saiatu atal honetan sortutako grafikoak errepikatzen.

EDUKIAK

Sarrera

Mapak

Datuak

Klusterrak

MAPAK



Mapak euskalgeotik²jaitsiko ditugu:

```
> eskualdeak.url <- "http://euskalgeo.net/sites/euskalgeo.net/
files/fitxategi-eranskin/Eskualdeak_0.zip"
```

Aurretik, jaitsitako artxiboa gordeko dugun karpeta sortuko dugu:

```
> euskalgeo <- "../datuak/euskalgeo/"
> dir.create(euskalgeo)
```

ZIP fitxategia jaitsi ostean, karpeta horretan erauziko dugu:

```
> eskualdeak.zip <- paste0(euskalgeo, "eskualdeak.zip")
> download.file(eskualdeak.url, eskualdeak.zip)
> unzip(eskualdeak.zip, exdir=paste0(euskalgeo, "eskualdeak"))
```

²<http://euskalgeo.eus/>

MAPAK

Mapak **shp** fitxategik irakurtzeko **rgdal**³ paketea erabiliko dugu:
'Geospatial' Data Abstraction Library ('GDAL')

```
> library(rgdal)
```

```
> eskualdeak.ftx <- "../datuak/euskalgeo/eskualdeak/  
    Eskualdeak.shp"  
> eskualdeak.shp <- readOGR(eskualdeak.ftx)
```

readOGR funtzioak **SpatialPolygonsDataFrame** klaseko objektu bat itzultzen du.

³<https://cran.r-project.org/web/packages/sp/sp.pdf>

³<https://cran.r-project.org/web/packages/rgdal/rgdal.pdf>

³<http://www.gdal.org/>

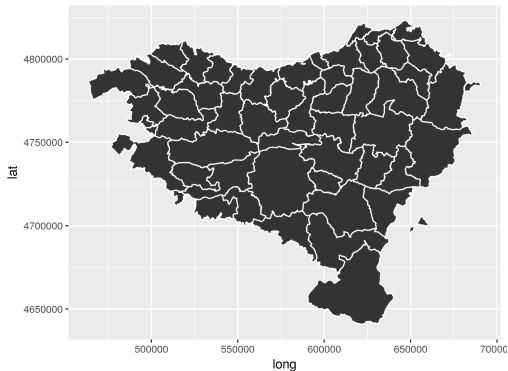
MAPAK

ggplot erabiliz mapa irudikatzeko, **SpatialPolygonsDataFrame** motatik **data.frame** motara itzuli behar da. Horretarako **broom**⁴ paketea erabiliko dugu:

```
> library(broom)
> eskualdeak <- tidy(eskualdeak.shp,region="es_kod_2")
> str(eskualdeak)
'data.frame': 276995 obs. of 7 variables:
 $ long      : num   545648 545643 545617 545563 ...
 $ lat       : num   4788984 4788976 4788846 4788733 ...
 $ order     : int     1  2  3  4  5  6  7  8  9 10 ...
 $ hole      : logi    FALSE FALSE FALSE FALSE ...
 $ piece     : Factor w/ 7 levels "1","2","3","4",...: 1 1 1 1 ...
 $ group     : Factor w/ 74 levels "0.1","0.2","0.3",...: 1 1 1 1 ...
 $ id        : chr     "0" "0" "0" "0" ...
```

⁴<https://cran.r-project.org/web/packages/broom/broom.pdf>

MAPAK



```
> g <- ggplot()
> g <- g + geom_map(data=eskualdeak      # Datuak
                    , map=eskualdeak    # Oinarria
                    , aes(x=long, y=lat, group=group, map_id=id)
                    , mcolor="white")   # Mugen kolorea
> ggsave("mapa.png", plot=g , width = 6, height = 4.5)
```

ARIKETAK II

1. Irudikatu Euskal Herriko udalerrien mapa.

EDUKIAK

Sarrera

Mapak

Datuak

Klusterrak

DATUAK



Open Data Euskadi

- ▶ Jaurlaritzaren eta bere menpeko erakundeen datu-irekien ataria.
- ▶ 2010ean egin zen publiko ataria.
- ▶ <http://opendata.euskadi.eus/hasiera/>

Ezaugarriak

- ▶ 4000tik gora datu-multzo.
- ▶ Datuak CSV, JSON eta XML formatuetan eskuragarri.
- ▶ SPARQL erabiliz datu-kataloan koltsultak egiteko aukera.

DATUAK

Fitxategia Editatu Ikusi Txertatu Formaturia Orria Datuak Tresnak Leihos Laguntza									
Xehetasuna eskualdearen arabera									
A	B	C	D	E	F	G	H	I	J
1	Udalerrietako iraunkortasun adierazleak	Nekazaritza eta arrantza sektorean okupatutako urte edo gehiagoko biztanleria (%)							
2									
3	Datuen laburpena								
4	Entitatea	2016	2015	2011	2010	2006	2001	1996	
5	Arabako Araba	2,09	2,63	1,86	1,58	1,48	2,77	3,70	
6	Bizkaia	0,91	1,30	0,86	0,77	0,94	1,51	1,93	
7	CAE	1,11	1,54	1,03	0,91	1,05	1,78	2,37	
8	Gipuzkoa	0,97	1,40	0,89	0,81	1,03	1,74	2,44	
9									
10	Xehetasuna eskualdearen arabera								
11	Lurralde kodea	Lurralde	2016	2015	2011	2010	2006	2001	1996
12	1100	Arabako Ibarrek / Valles Alaveses	10,39	11,41	8,56	6,87	8,06	17,45	30,90
13	1200	Arabako Lautada / Llanada Alavesa	0,92	1,43	0,88	0,76	0,75	1,26	1,47
14	1300	Arabako Mendialdea / Montaña Alavesa	12,16	13,31	9,84	10,33	12,13	21,80	30,30
15	48100	Arratia Nerbioi / Arratia-Nervi	2,55	2,88	2,17	1,73	1,84	2,78	3,76
16	20200	Bidasoa Beherea / Bajo Bidasoa	0,90	1,57	1,04	0,96	1,37	2,31	3,15
17	48200	Bilbo Handia / Gran Bilbao	0,41	0,82	0,39	0,36	0,43	0,51	0,42
18	20100	Deba Beherea / Bajo Deba	1,25	1,73	1,16	1,03	1,02	1,69	2,76
19	20300	Debagoiena / Alto Deba	0,67	0,96	0,62	0,57	0,56	0,83	1,06
20	20400	Donostialdea / Donostia-San Sebastián	0,59	1,00	0,56	0,54	0,73	1,28	1,65
21	48300	Durangaldea / Duranguesado	1,03	1,34	0,88	0,82	0,82	1,14	1,12
22	48400	Enkartzazioak / Encartaciones	4,16	4,76	3,76	3,04	3,54	6,73	9,21
23	1400	Emioxa Arabarra / Rioja Alavesa	21,87	22,06	17,92	14,45	10,83	21,91	28,21
24	48500	Gernika-Bermeo	3,34	3,85	3,31	2,97	4,63	9,38	13,77
25	20500	Goierri	1,11	1,50	0,98	0,87	1,02	1,43	2,18
26	1500	Gorbeia Inguruak / Estribaciones del Gorbea	4,51	4,83	4,22	3,44	3,71	6,48	10,53
27	1600	Kantauri Arabarra / Cantabria Alavesa	2,04	2,58	1,83	1,64	1,54	3,33	3,92
28	48600	Markina-Ondarroa	5,54	5,75	6,37	6,03	7,64	13,03	19,23
29	48700	Plentzia-Mungia	1,62	1,87	1,37	1,19	1,32	2,43	3,71
30	20600	Tolosaldea / Tolosa	1,76	2,20	1,37	1,22	1,35	2,43	3,79
31	20700	Urola-Kostaldea / Urola Costa	2,09	2,46	1,91	1,63	2,26	4,02	5,92
32									
33	Xehetasuna udalerriaren arabera								
34	Udalerri kodea	Udalerria	2016	2015	2011	2010	2006	2001	1996
35	48001	Abadiño	1,10	1,44	1,15	1,29	1,57	1,78	0,97
36	20001	Abejuelo	6,62	7,88	5,26	3,47	6,20	11,28	18,10

indicator-1



Bilatu



Bilatu denak



Formatudun bistaratzeko



Bereizi maiuskula/minuskulak



Lehenetsia

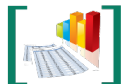
Batez bestekoa: 0 / Batuketak: 0

% 140

DATUAK



Gaindegia



datuak.net

Euskal Herriko *datu biltegia*

Datuak.net

- ▶ Euskal Herriari buruzko datu fitxategiak modu irekian eskaintzen dituen gunea.
- ▶ Gaindegiaren plataforma.
- ▶ <http://datuak.eus>

Ezaugarriak


- ▶ 47 datu multzo.
- ▶ Datuak ODS, XLSX formatuetan eskuragarri.
- ▶ CKAN API-a erabiliz datu-kataloan koltsultak egiteko aukera.

DATUAK

Fitza batzuetan editatu ikusi Tseriatu formatua Orria Datuak Itxuratu Lehenoa Laguntza

Calibri 11

K3

 datuak.net Euskal Herriko datu biltegia				www-datuak-net EUSKAL HERRIKO DATU BILTEGIA						
				www.gaindegia.org		www.atlasa.net		www.euskalgeo.net		
Biztanleria adin-talde nagusien arabera. Euskal Herria eta bere udalerriak, 2016.										
Hurrialde kodea	Herrialdea	Euskarazko izena	Izen ofiziala	Orotara	0-14	15-64	65 eta +	0-14 (%)	15-64 (%)	65 eta + (%)
BH		EUSKAL HERRIA	Euskal Herria / P ^z	3.133.122	455.018	2.015.976	662.127	14,5	64,3	21,1
48001	Bizkaja	Abadiño	Abadiño	7.533	1.268	4.951	1.314	16,8	65,7	17,4
31001	Nafarroa Garaia	Abajgar	Abajgar	90	4	63	23	4,4	70,0	25,6
20001	Gipuzkoa	Abaltzisketa	Abaltzisketa	324	59	208	57	18,2	64,2	17,6
48002	Bizkaja	Abanto	Abanto y Ciérvano	9.577	1.337	6.615	1.625	14,0	69,1	17,0
31002	Nafarroa Garaia	Abartzuzza	Abartzuzza <> Abar	532	73	308	151	13,7	57,9	28,4
31003	Nafarroa Garaia	Abaurregaina	Abaurregaina/Ab	127	7	71	49	5,5	55,9	38,6
31004	Nafarroa Garaia	Abaurrepea	Abaurrepea/Abau	35	0	23	12	0,0	65,7	34,3
31005	Nafarroa Garaia	Aberin	Aberin	376	36	255	85	9,6	67,8	22,6
31006	Nafarroa Garaia	Abiltas	Abiltas	2.494	354	1.560	580	14,2	62,6	23,3
31007	Nafarroa Garaia	Adios	Adios	155	22	98	35	14,2	63,2	22,6
20002	Gipuzkoa	Aduna	Aduna	470	99	304	67	21,1	64,7	14,3
31019	Nafarroa Garaia	Agoitz	Aoiz/Agoitz	2.564	449	1.675	440	17,5	65,3	17,2

data metadata

Bilatu Bilatu denak Formatuakun bistaratzeko Bereizi maiuskula/minuskulak

1 / 2 orria PageStyle_data Baitez bestekoa : Batuketa: 0 % 160

DATUAK

R bidez Gaindegiaiko datuen katalogoa ikusteko **ckanr**⁵erabiliko dugu:

```
> library(ckanr)
> ckanr_setup(url = "http://datuak.eus")
> datu_multzoak <- package_list(as = "table")
> head(datu_multzoak)
[ 1 ] "autopista-sarearen-dentsitatea"
[ 2 ] "barne-produktu-gordina"
[ 3 ] "berrikuntzako-adierazleen-panela"
[ 4 ] "biztanleria-bost-urteko-adin-taldeen-arabera"
[ 5 ] "biztanleria-dentsitatea"
[ 6 ] "biztanleria-hezkuntza-mailaren-arabera"
```

⁵<https://cran.r-project.org/web/packages/ckanr/ckanr.pdf>

DATUAK

```
> multzoa <- package_show(datu_multzoak[4])
> multzoa
<CKAN Package> 8de97c70-280c-4750-93bd-39b0094397f1
Title: Biztanleria bost urteko adin-taldeen arabera
Creator/Modified: 2013-12-23T11:31:31.388856 /
                2017-10-12T11:23:13.741843
Resources (up to 5): Sexua eta bost urteko adin taldeen arabera
Tags (up to 5): adin-taldea, adin-tartea, adina, biztanleak, ...
Groups (up to 5): biztanleria
```

```
> multzoa$name # izena ikusteko
[1] biztanleria-bost-urteko-adin-taldeen-arabera
> multzoa$resources # baliabideak
> length(multzoa$resources)
[1] 16
> baliabidea <- multzoa$resources[[7]]
```

```
> adierazlea.fitx <- paste0(multzoa$name , '.'
                             , tolower(baliabidea$format))
> download.file(baliabidea$url, adierazlea.fitx)
```

```
> library(readxl)
> adierazlea.taula <- read_excel(adierazlea.fitx
                                , skip=10,col_types="text")
```

```
> adierazlea.taula
```

```
# A tibble: 686 x 41
```

	'Lurralde kodea'		Herrialdea		'Euskarazko izena'
	<chr>		<chr>		<chr>
1	EH		NA		EUSKAL HERRIA
2	48001		Bizkaia		Abadiño
3	31001	Nafarroa Garaia			Abaigar
4	20001		Gipuzkoa		Abaltzisketa
5	48002		Bizkaia		Abanto
6	31002	Nafarroa Garaia			Abartzuza

DATUAK

tidyr⁶ eta **dplyr**⁷ datu-taulak aldatu, garbitu eta batzeko bi pakete eraginkor dira.

```
> library(tidyr)
> library(dplyr)
> adinka <- adierazlea.taula %>%
  filter('Lurralde kodea' != "EH") %>%
  gather(adina, balioa, 5:ncol(adierazlea.taula)) %>%
  mutate(balioa = as.numeric(balioa))
```

Aukeratutako adierazlean, era absolutuan eta ehunekoetan dago adierazita biztanleria adin tarte bakoitzeko . Guk ehunekoak erabiliko ditugu.

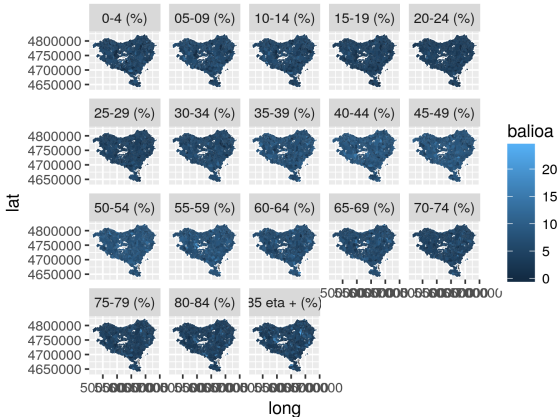
```
> adinka.ehun <- adinka %>%
  filter(grepl("%",adina))
```

⁵<https://cran.r-project.org/web/packages/tidyr/tidyr.pdf>

⁷<https://cran.r-project.org/web/packages/dplyr/dplyr.pdf>

⁷www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

DATUAK



```

> g <- g + geom_map(data=adinka.ehun, map=udalerrriak,
                    aes(map_id='Lurralde kodea',
                        color=factor(adina), fill=balioa),
                    color="#7f7f7f", size=0.01)
> g <- g + facet_wrap(~ adina) # Adin tarte bakoitza grafiko bat
> g

```

ARIKETAK III

1. Sortu hiru aldagai berri "0-24", "25-69" eta "70 eta +". Hauek, herri bakoitzean, adin tarte bakoitzerako dagoen biztanle kopurua adieraziko dute.
2. Kalkulatu, herri bakoitzerako, tarte bakoitzaren kopuru erlatiboa.
3. Errepikatu azken grafikoa, "0-24", "25-69" eta "70 eta +" adin tarteak erabiliz.

EDUKIAK

Sarrera

Mapak

Datuak

Klusterrak

KLUSTERRAK

Instantziak eta aldagaiak taldekatzeko hainbat pakete daude R-n. Oinarrizkoak, eta ezagunenak, hclust eta kmeans dira.

- ▶ hclust: Kluster hierarkikoak sortzen ditu. Datuen arteko distanzian oinarrituz, pauso bakoitzean bi talde batzen ditu.
- ▶ kmeans: Datuen partizio batetatik abiatuz, pausu bakoitzean tamaina berdineko partizio bat sortzen du zentroideetarako distantzian oinarrituz.

Funtzio hauek aplikatzeko datuak koordinatu eran adierazi behar ditugu. Datuak eskalatzea eta zentratzea komeni da, algoritmoen konbergentzian arazoak gutxitzeko.

```
> adinka.ehun2 <- adinka.ehun %>%  
  spread( adina, balioa ) %>%  
  select( - Herrialdea, -'Euskarazko izena',  
         - 'Izen ofiziala' ) %>%  
  scale()
```

KMEANS

```
> biztanleria.tald <- kmeans(adinka.ehun2[, -1], centers = 3)
```

```
> biztanleria.tald
```

K-means clustering with 3 clusters of sizes 301, 302, 82

Clusters means:

	0-4 (%)	05-09 (%)	10-14 (%)	15-19 (%)	20-24 (%)
1	3.9391	4.6993	4.7508	4.1770	3.9112
2	5.8246	6.4719	5.8280	4.5662	3.8901
3	1.7172	2.5756	2.0489	2.1873	2.8994

Clusters vector:

```
[ 1 ] 2 1 2 2 1 3 3 1 1 1 2 2 3 2 1 2 2 2 2 2 1 1 2 2 2 1 1 2 2 1 2 2 2 1 1
[ 37 ] 2 1 2 2 1 2 1 1 1 2 1 2 1 3 2 1 1 2 2 2 2 2 2 1 2 1 2 1 1 2 1 1 2 3
```

Within cluster sum of squares by cluster:

```
[ 1 ] 12333.3 14325.9 8613.1
```

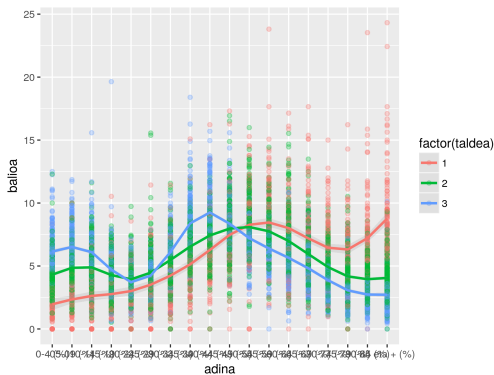
(between_SS / total_SS = 26.6 %)

Available components:

```
[ 1 ] "cluster"      "centers"      "totss"      "withinss"    "tot.withinss"
[ 2 ] "betweenss"    "size"        "iter"      "ifault"
```

```
> adinka.ehun2$aldeia <- biztanleria.tald$cluster
```

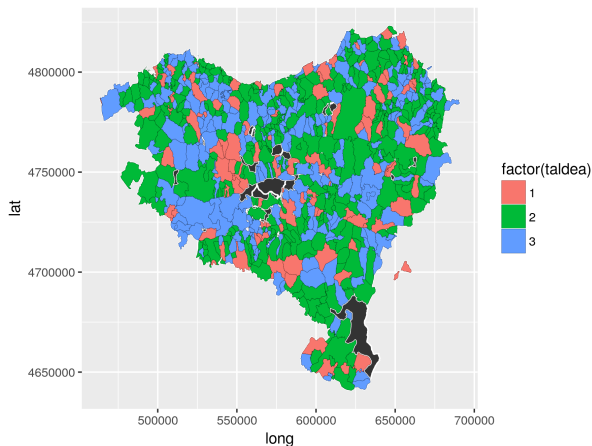

KMEANS



```

> adinka.ehun3 <- adinka.ehun2 %>%
  gather(adina, balioa, 2:19)
> g <- ggplot(adinka.ehun3,
  aes(x=adina, y=balioa,
  colour=factor(taldea), group=taldea))
> g <- g + geom_point(alpha=.3) +
> g <- g + geom_smooth(alpha=.2, size=1)

```



```
> g <- g + geom_map(data=adinka.ehun2, map=udalerrriak,  
  aes(map_id='Lurralde kodea', fill=factor(taldea)),  
  color="black", size=0.05)
```

kmeans metodoarekin eta kluster hierarkikoekin parametro batzuk zehaztu behar ditugu:

- ▶ **kmeans**-en talde kopurua
- ▶ **hclust**-en talde kopurua edo bi talde batzeko irizpidea

Parametro hauen balio optimoa aukeratzeko, pakete gehigarriak aurkitu ditzakegu. Adibidez, **kmeans**entzat talde kopuru egokiena aukeratzeko **fpc**⁸ paketea erabili dezakegu.

```
> library(fpc)
> biztanleria.tald2 <- pamk( adierazlea.ehun2[,-1],
                           krange=1:5, critout=TRUE)
1 clusters 0
2 clusters 0.1549
3 clusters 0.061215
4 clusters 0.036129
5 clusters 0.030447
> adinka.ehun2$alde2 <- biztanleria.tald2$pamobject$clustering
```

⁸<https://cran.r-project.org/web/packages/fpc/fpc.pdf>

ARIKETAK IV

1. Taldekatu udalerriak biztanleriaren arabera 2 taldetan.
2. Aukeratu beste datu multzo bat (udalerrientzako), eta kalkulatu honentzako oinarrizko neurri estatistikoak. Egin hauen adierazpen grafikoa ggplot erabiliz.
3. Gurutzatu bi datu-multzoak hurrengo zentzuan: aukeratutako datu multzoari gehitu aldagai berri bat, herri bakoitzaren taldea adierazten duena (?full_join).
4. Grafikoki aztertu datu multzo berriaren oinarrizko neurri estatistikoen portaera bi talde desberdinen arabera.

Eskerrik asko!



ueu
udako
euskal unibertsitatea



gkobeaga@bcamath.org