eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

EKONOMIA ETA ENPRESA FAKULTATEA

FACULTAD DE ECONOMÍA Y EMPRESA

HISTOGRAMEN ERAKETA R SOFTWAREAN

Ekonomia eta Enpresa Fakultatea – Gipuzkoa

Gradu amaierako lana

Aitor Urretabizkaia Olea

Zuzendaria: Jose Mari Sarasola

Defentsa urtea: 2020

Aurkibidea

1 SARRERA 1
2 HISTOGRAMAK 4
2.1 Histogramen Definizioa eta historia 4
2.2 Histograma baten eraketa 6
2.2.1 Tarteak definitu 7
3 R Project programa informatikoa 20
3.1 Garapen historikoa 20
3.2 Nork erabiltzen du R? 22
3.3 R *softwareak* dauzkan abantaila eta desabantailak 27
3.3.1 R *softwareak* eskainitako abantailak 27
3.3.2 R softwarearen desabantailak 30
3.4 R softwarea MatLabekin konparatuta 30
3.5 R Commander eta R Studio 33
4 R SOFTWAREAN HISTOGRAMAK ERATZEKO GIDALIBURUA 36
4.1 DATUAK BILTZEA 36
4.2 Read R paketea 44
4.3 Datuen prestakuntza 44
4.4 TAULARAKETA ETA GRAFIKOAREN SORRERA 46
4.4.1 Psych paketea 50
4.4.2 Lattice paketea 53
4.4.3 GGPLOT2 paketea 56
4.4.4 Patchwork paketea 59
4.4.5 Agricolae paketea 62
4.5 HISTOGRAMAREN EDIZIOA 65
4.5.1 KOLOREAK 65
4.5.2 Lerroak 70
5 ONDORIOAK 72
6 BIBLIOGRAFIA 75

1 SARRERA

Zenbat eta teknologia gehiago eskura izan, enpresek, datu gehiago biltzeko eta ikertzeko aukerak dituzte. Batzuek, etekin handiagoak izango dituzte, eta beste batzuek, murritzagoak. Baino nabarmentzen dena, hauxe da: lortu daitezkeen datuak ondo ikertu ondoren, enpresak bere helburuak betetzeko izango dituen aukerak handiagoak izango direla. Adibide bezala, gaur egun liderrak diren enpresa dauzkagu; *Google, Facebook, Twitter,*…enpresa hauek guztiak, beraiek biltzen dituzten datuetatik etekinak ateratzen dituzte. Datu pertsonalen salmenta gauza normala da gaur egun, irabazi handiak sortuz.

Lan honetan, datuak tratatzeko tresna desberdinak ikusiko ditugu. Alde batetik, grafikoak oso baliagarriak izango zaizkigu, jasotako informazio guztia biltzeko eta era egokian tratatzeko edota aurkezteko. Grafikoak, milaka datu irudikatzeko aukera emango digute, irudi batekin. Horrela, jasotako informazio guztia era errazago batean barneratuko dugu. Lan honen helburuetako bat hau da; histogramek informazio tratamenduan izan dezaketen garrantzia ikertzea. Beste grafiko batzuen alde izango dituen abantailak ikusiko ditugu.

Beraz, lan honen helburuak hauek izango dira. Lehen aipatu dugun bezala, histogramen garrantzia ulertzea eta histogramen eraketa ezagutzea izango da beste batzuen artean. Gainera, lan hau aprobetxatu nahi dugu, R programak eskaini ditzakeen pakete ezberdinak aztertu eta aurkezteko. Era honetan, programa honekin egin ditzakegun aukera ezberdinak ikusiko ditugu, irakurlearentzat gida bat eraikiz. Behin gida bat sortu dugula, eta R programa informatikoarekin lan egin dugula, aztertutako guztia laburbilduko dugu ondorioak bilatuz.

Aipatu dugun bezala, R programarekin lan egingo dugu, histogramekin lan eginez. Bi elementu hauek, estatistikan irakasten diren elementuak izango dira. R programak datuak gordetzeko eta tratatzeko aukera ezberdin asko eskainiko dizkigu, hauen artean grafikoak sortzekoa aurrerago ikusiko dugun bezala. Mundu mailan, asko erabiltzen den programa da, eta *Big Data*

(ondoren aztertuko dugun kontzeptua) mundua enpresa garapenean izugarri handitzen ari da. Beraz, programaren ezagutza ekonomialari bezala lagungarria izan daiteke lan munduan hastean.

Hau izan daiteke, nik lan hau aukeratzeko izan dudan motibazioetako bat. Iragan profesionalean garrantzitsua izan daitekeen programa izan daiteke. Bereziki gustuko liratekeen lana, banku munduan izango zen. Bertan, banketxe batzuetan erabiltzen jakitea eskatzen duten programa R izango da. Honekin, banketxe ezberdineko ateak irekita izatea lortu dezaket. Adibidez, Santander Banketxeak ezinbesteko ezaugarri bezala eskatzen du, bere langile berrien artean, R programa erabiltzen jakitea, ondoren ikusiko dugun bezala.

Honetaz aparte, lan hau garatzeko, interes handi bat beharrezkoa da. Horregatik pentsatu nuen estatistikarekin zerikusia izan zezakeen zerbait lantzea, ikasketan gustukoenen izan duten ikasgai bat izan da, eta edozein momentutan erabilgarri izan daitekeena gainera, ez bakarrik enpresa munduan (beste batzuk ez bezala). Gainera, finka administrazioaren barnean lan egiten dut eta R programa erabilgarria izan daiteke kalkulu batzuk egiteko, aurkezpenak egiteko,…hau da, Bizilagunentzat informazioa era erosoago batean emateko.

Lan honetan, metodologia ezberdinak erabili ditut emaitzak lortzeko. Aipagarriena, Internetek eskaini didan laguntza izan da. R programak Interneten gidaliburu ezberdinak eskaintzen ditu, guztiz publikoak eta fidagarritasun handia eskaintzen dutenak. Nahiz eta Interneten informazio asko izan, EHUko liburutegi zerbitzuak laguntza konplementario bat eskaini dit, honek dauzkan liburuetan informazioa bilatzeko aukera eskainiz.

Esan daiteke informazioa bilatzeko problemarik ez dutela izan, programa irekia denez, jendeak asko lantzen duen *softwarea* baita, eta informazioa partekatzeko ohitura dago, *GNU* komunitatearen artean. Baliagarria izan da baita ere, estatistika eta enpresa munduan garrantzia duten organismoekin lan egitea. Adibidez, *EUSTAT*etik informazio baliagarri asko lortzeko aukera izan dut. Beraz, gure lanaren egitura honela definituko dugu.

Lehenengo, histogramaren kontzeptua ezagutu beharko dugu. Histograma bat nola eratzen den ulertu eta mota ezberdinak aztertu beharko ditugu. Histogramen eraketa nola burutzen den ulertu ondoren, R programa ezagutuko dugu. Programa honek dauzkan abantailak aipatuko ditugu. Programak erabiltzen duen hizkuntza (S) ezagutuko dugu eta honek izan duen garapen historikoa aipatuko dugu. Zati hau bukatzeko, bere konpetentzia diren programa ezberdinekin alderatuko dugu eta R pakete ezberdinak aipatuko ditugu, erabiltzaile basiko batentzat laguntza bezala izan daitezkeenak programazio noziorik ez baldin badu.

Hau guztia aztertu ondoren, R *softwarea* erabiltzen hasiko gara. Hasieran kontzeptu basiko batzuk azalduko ditugu. Adibidez, datuak nola barneratzea irakatsiko dugu. Honekin bukatzean, R programak eskaintzen dizkigun pakete ezberdinak aztertuko ditugu. Pakete hauek, histogramen eraketan bereziki erabili daitezkeen paketeak izango dira, eta bertan erabilgarriak diren komando aipagarrienak definituko ditugu. Era honetan, programazio ezagutzarik ez duen erabiltzaile batek, lan honen laguntzarekin, grafiko ezberdinak eratzeko gai izango da.

Gida honekin bukatzeko, gure histogramaren aurkezpenean lagundu ditzakeen faktoreak aldatzea ikasiko dugu. Koloreak, lerroen lodiera,…Baita ere oso garrantzitsua izango da, hemendik lortutako emaitzak era egokian gordetzen jakitea (*PDF,JPG*,…). Gida bukatzeko, R kontsolarekin burututako grafikoak era egokian gordetzen erakutsiko dugu. Hau burutu hondoren, irakatsitako guztia praktikan jarriko dugu, kasu praktiko labur batekin, ikusteko R programarekin burutu daitezkeen ekintza ezberdinak, kasu erreal batean.

Hau burutu einean, lan honetatik lortu ditugun ideia nagusiak bilduko ditugu, eta ikasitako guztia laburbilduko dugu. Lanarekin hasi baino lehen, gustatuko litzaidake tutoreari egindako lan guztia eta eskainitako laguntza guztia eskertzea, eta unibertsitateari, eskainitako baliabideak eskaintzeagatik. Besterik gabe, has gaitezen histogramak aztertzen eta hauen eraketarako garrantzitsuak diren faktoreak aztertzen.

2 HISTOGRAMAK

2.1 Histogramen Definizioa eta historia

Histograma bat, barra diagrama mota bat da, eta datu bat edo batzuk tarte batean zenbat alditan errepikatzen den irudikatuko digu, hau da, maiztasuna irudikatzen duen grafika izango da. Ezagutzen den lehen histograma, *Kooi Rren PhD tesian* agertu ziren[1] 1880 urtean, baina 1985 urte arte, histograma hitzarekin ez zen ezagutzen. Karl Pearson[2] estatistikari famatuak izen hau eman zion arte.

Oxford hiztegi ingelesean diotenez, *"Philosophical Transactions of the Royal Society of London" Series A, Vol. CLXXXVI, (1895) p. 399, it is mentioned that "[The word 'histogram' was] introduced by the writer in his lectures on statistics as a term for a common form of graphical representation, i.e., by columns marking as areas the frequency corresponding to the range of their base."* Idazleak izen hau jarri zion lehenagotik ikusten zen grafiko bati. Beraz, hauen lehenengo eredua ez dago argi definituta non azaldu zen. Dakigun guztia, histogramak barra diagramak eratzeko beste modu bat direla eta hauek lehen aldiz 1786 urtean[3] publikatu zirela Williem Playfair-ek eratutako liburutik. Estatistikari honek, barra diagramekin batera, sektore diagramak eratu zituen. Estatika deskribatzailearen matematikari garrantzitsuenetako bat bezala mintzatzen da Espainiako Estatistika Institutuan[4].

XX Mendean, informatika garapenarekin batera, data base informatikoak sortzen dira. Honekin, informazioa gordetzeko eta erabiltzeko mundu berri bat irekitzen da. Mende honetan, errepresentazio grafikoak bultzada izango du ere bai. Pertsona batek ezin ditu milioi bat datu interpretatu, grafikoki ipinita ez badaude. Beraz, estatistika deskribatzaileak datuak jasotzeko, ordenatzeko eta klasifikatzeko behar zuen bultzada jaso zuen, gaur egun ezagutzen duguna arte. "*Estatistika deskribatzailea lagin*

[1] Kooi R. (1980). The Optimization of Queries in Relational Databases. Cleveland: PhD Thesis Case Western Reserve University
[2] Bere ekarpenen artean, desbiderapen estandarra eta chi karratuaren froga estatistikoa dauzkagu
[3] Playfaire, W. (1786). *The Commercial and Political Atlas*. London: Cambridge University Press
[4] INE. (2020). *Dos siglos de gráficos estadísticos: 1750 - 1950 / 1801 - 1850 / William Playfair (1759-1823).* Ine.es. Iturria: https://www.ine.es/expo_graficos2010/expogra_autor2.htm.

bateko informazioa laburbiltzen duten grafikoak (histograma, sektore-diagramak) eta neurri estatistikoak (batezbestekoa, desbideratze estandarra, ...) aurkeztu egiten dituen estatistikaren adarra da, lagin horretako emaitzak populaziora zabaldu gabe, eta beraz haietan egin daitekeen lagin errorea kontuan hartu gabe." (Gizapedia)

Grafiko mota hauek, datuen ulermen azkarragoa lortzea dute helburu, irakurlearen interpretazioak zehatzagoak eta azkarragoak izateko. Zehatzago esanda, histogramek datu kopuru bat zenbat alditan errepikatzen den X ardatzean kokatutako serieetan irudikatuko digu.

Askotan, datuak biltzeko eredu hau barra diagrama batekin nahastu daiteke. Histograma bat identifikatzeko, datuetan fijatu behar gara. Espainiako Estatistikako Institutuak[5], histogramak eratzeko behar diren datuak aldagai kuantitatibo jarraiak izan beharko direla zehazten du.

Histograma batek errepresentatzen duena ulertzeko, jo dezagun produktu berri bat merkaturatu nahi dugula. Horretarako, produktu honen posizionamendu estrategikoa biztanleriaren altuerarekin erlazionatuko dugu. Beraz, saiatu behar gara produktu hau, biztanleria kopuru nabarmen bati iristea. Horretarako, histograma bat eratu beharko dugu. Modu honetan, gure ikerketan gehien errepikatzen diren altuerak definitu ahal izango ditugu. Ikus dezagun, beraz, histograma baten eraketa nola burutu, baino lehenago, ikus dezagun zein kasutan eratu behar dugun histograma, eta zein kasutan zutabe diagrama. Horretarako, gizapedia webguneak horrela desberdinduko ditu:

"Estatistikan, aldagai estatistikoen artean aldagai diskretuak eta aldagai jarraituak (edo jarraiak) bereizten dira. Aldagai diskretuak balio isolatuak, banan-banan konta daitezkeenak, hartzen dituen aldagaiak dira, hala nola familiako anai arreba kopurua (0, 1, 2, 3, ...). Aldagai jarraituak berriz, balio ezberdin asko hartzen dituzten aldagaiak dira, edota tarte batean zehar edozein balio har dezaketenak; adibidez, pertsona baten altuera aldagai jarraitua da, balio desberdin asko har ditzakeelako (150, 151,

[5] Instituto Nacional de Estadistica. (2015). *Tipos de gráficos*. Ine.es. Iturria: https://www.ine.es/explica/docs/pasos_tipos_graficos.pdf.INEren (Instituto Nacional de Estadistica) txostenean agertzen diren grafika moten ezaugarriak azaltzen dira. Bertan, histograma batek bete behar dituen ezaugarriak azaltzen dira.

152, ..., 210 cm), edo baita ere 150-210 tartean (edo tarte zabalago batean) edozein balio.

*Aldagai diskretu eta jarraituen arteko bereizketa garrantzitsua datuak nola modelizatu eta datuei aplikatu beharreko teknika estatistikoak erabakitzean. Adibidez, datu diskretuak eta jarraituak grafikoki irudikatzeko diagramak ezberdinak dira: aldagai diskretuetarako, zutabe-diagrama erabiltzen den bitartean, aldagai jarraituetarako histograma erabiltzen da."(*Gizapedia[6])

Beraz, histograma bat eratzen hasi baino leen datuetan erreparatu behar dugu. Behin datu hauek aldagai jarraituak direla baieztatu dugu, histograma eratzen hasiko gara.

2.2 Histograma baten eraketa

Histograma bat, lehen esan dugun bezala, datuen maiztasuna irudikatzen duen grafika izango da. Grafika hau burutzeko, abiapuntu batzuk jarraitu beharko ditugu datuak era egokian ulertzeko. Beraz, histograma bat zer den ulertzeko, imagina dezagun 20 pertsonen altuerak dauzkagula. Nahiko datu dauzkagunez, histograma sinple bat eratuko dugu eta eredu bakoitzerako zenbat tarte izango zituen ereduak aipatuko dugu. Hau egin ondoren, eredu guztiak konparatuko ditugu.

Histograma eratzen hasi baino lehen, aipagarria izango da, zenbat eta tarte gutxiago, geroz eta informazio gutxiago jasoko dugula. Tarte gehiegi egoteak ere kalte egingo digu, beraz, datuak irudikatzeko eredu optimoa bilatzeko, matematikariek eredu desberdinak proposatu dituzte historian zehar. Ez dago eredu matematiko definitiborik, zenbat tarte erabili behar diren definituko dituena. Hurrengo ereduetan, proposatuko ditugun metodoak aipatuko ditugu eta honekin batera, método hauek aplikatzea gomendatzen diren ereduak ikusiko ditugu.

[6] Sarasola, J.M. (2020ean egiaztatuta). Aldagai diskretu eta aldagai jarraituak. Gizapedia.hirusta.io. Iturria: https://gizapedia.hirusta.io/aldagai-diskretuak-eta-aldagai-jarraituak-jarraiak/.

Oso arrunta da, ikertzaile edo pertsonen artean, datuak era egokian biltzea, eta informazio guztiarekin, ondorio okerrak biltzea. Kasu hau, aukera ezberdinengatik gertatu daiteke. Hauen artean, histogramen kasuan, tarteen definizioa izango da. Nahiz eta oso garrantzitsua ez diruditu, kontrakoa izango da, tarteak era egokian eratzea ezinbestekoa izango da. Horretarako, matematikak laguntza eskeiniko digu, erregela ezberdinak proposatuz gure datu kopurura moldeatzeko. Ikus ditzagun, tarteak definitzeko erregela ezagunenak zeintzuk diren eta nola aplikatu daitezkeen

2.2.1 Tarteak definitu

Sturges Erregela:

Nahiz eta onartutako hipotesiak oso murritzak izan eta horregatik oinarri estatistiko eskasa izan, oso maiz erabiltzen den metodoa da Sturges erregela. Formula honek eskaintzen digun erreztasuna, datu kopuru gutxi dauzkagunean izango da (*N* 200 baino baxuagoa denean[7]).

Datu kopurua aipatutako tamaina baino murritzagoa baldin bada, konplexutasun haundiagoa duten formulen antzeko emaizak lortuko ditugu. Metodo hau, 200 datu baino gehiagoko datuetan aplikatuko bagenu, histograma leunduko luke, bereziki, alborapen haundia eta moda anitzak dauzkaten datuak erabiltzen badira.

Beraz, laburbilduz, *n* tarte kopuruaren arabera, honako formula hau aplikatuko dugu tarte ezberdinak definitzeko:

$$K= 1+\log_2 n$$

Laburbilduz, honako taula hau eratu dezakegu n aukera guztietarako (*n*<200) k zenbatekoa izango zen.

[7] Estatistika nomenklaturan, askotan n eta k letrak erabiltzen ditugu, datu kopurua zenbatekoa den adierazteko. Lan honetan, nomenklatura hauek erabiliko ditugu.

N (DATU KOPURUA)	*K* (TARTE KOPURUA)
20-32	6
32-64	7
64-128	8
128-200	9

Taula 1Tarte kopurua n aukeren arabera

Dixon eta kronmalen erregela:

Sturgesen erregelaren antzekoa izango da, baino zehaztasun gutxiagokoa datu asko erabiltzen direnean. Kasu honetan, komenigarria izango da datuak prozesatzeko método hau erabiltzean, datu kopurua 100 baino gutxiagokoa izatea. Formula, honako hau izango da. $K=1+[\log_{10}(n)]$

Erregela enpirikoa edo Vellemanen erregela:

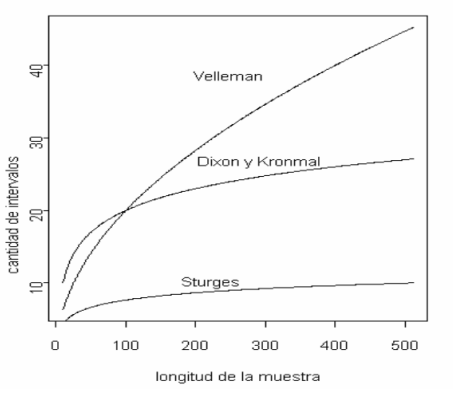
Askotan erabiltzen den erregela izango da. *Excel* programak automatikoki erabiliko duen erregela izango da, besteak beste.

Rice metodoa:

Aipatutako 3 metodoen bariantea izango da. Formula hau jarraituko du: $k= 2n^{1/3}$

Grafiko honetan[8], aipatutako 3 metodoek sortzen dituzten tarte kopurua azalduko da. Gogora ezagun, formula hauek datu kopuru gutxi daudenean erabili behar direla.

Irudia 1Tarteen arteko desberdintasunak. Buenos Aireseko Unibertsitateak egina



[8] Kelmansky, D. (2006) Departamento de Matemática de la facultad de ciencias exactas de Buenos Aires http://www.dm.uba.ar/materias/analisis_de_datos/2006/1/teoricas/Teor2a.pdf

Scotten erregela:

Histogramaren dentsitatea neurtzea ez da zaila izango *f(x)* formula izango duen dentsitate funtzio batetik. Tarte berdinak direla kontuan hartuta eta banaketa normal bat jarraitzen dutela kontuan hartuta $B(n, p_k)$, probabilitatearekin:

$$P_K = \int_{Bk} f(t)dt = \int_{tk}^{tk+h} f(t)dt.$$

Formula hau erabiliz, *Scott* estatistikariak, beste parámetro batzuen pean, formula hauek lortuko ditu:

$$h_s = (\frac{6}{n\int_{-\infty}^{\infty} f'(t)^2\,dt})^{1/3} \int_{-\infty}^{\infty} f'(t)^2\,dt = \frac{1}{4\sqrt{\pi\sigma^3}}$$

$$h_s = (\frac{24\sqrt{\pi\sigma^3}}{n})^{1/3} \approx 3{,}5\sigma n^{-1/3}$$

Beraz, tarteak kalkulatzeko formula hau izango da: (*s* desbiderazio standarra) (*n* tarteak).

$$h=\frac{3{,}49s}{n^{1/3}}$$

Freedman-Diaconis erregela:

David A. Freeman eta *Persi Diaconisek* sortutako teoría honek[9] tarte kopuruek kalkuluetan sortu dezaketen erroreen minimizazioa bilatuko du. Honako formula hau erabiliko dute, non *N* datu kopurua eta *IQ* kuartilen arteko ibiltartea:

$$h=2\frac{\mathrm{IQR(x)}}{\mathrm{n}^{1/3}}$$

[9] *Freedman, D.* eta *Diaconis, P.* (Abendua, 1981). *"On the histogram as a density estimator: L₂ theory".* Alemania: Springer Verlag. DOI 57 (4): 453–476.

Doaneren erregela:

Sturgesen erregelaren deribazio bat da, eta *Sturgesen erregela* bezela, ez da erabilgarria izango, datu kopuru haundia aztertu behar denean.

$$k = 1 + log_2(N) + log_2(1 + \frac{g_1}{\sigma_{g_1}})$$

Beste erregela erabili batzuk:

Oso erabilia den aukera bat, *n* datu kopuruen erro karratua izango da. Ez dauka inongo arrazoi matematikorik, baino datu gutxi dauden ereduetan, aplikagarria izango da.

Baita ere, oinarri teorikorik gabekoa, baino batzuetan erabilgarria izan daitekeen formula (datu kopuru gutxi dauden ereduetan) $k=N^{1/2}$ izango da.

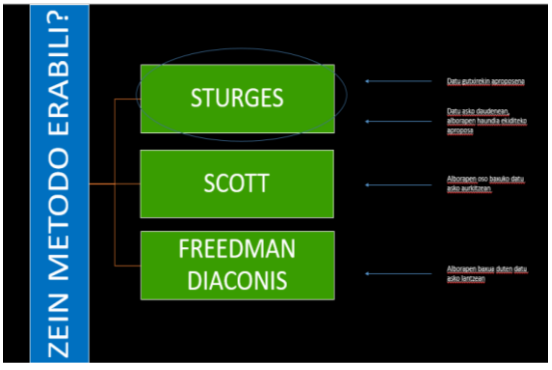
Eredu hauek guztiak kontuan izanda, konklusio hauek atera ditzakegu:

- Datu kopuru gutxirekin lan egiten dugunean, *Sturgesen metodoarekin* eratutako tarte kopuruak gomendatzen diren tarte kopuruaren (5-20) barne egongo dira. Beste aukera bat aukeratu ezkero, gomendatutako tarte kopuru minimoa baino gutxiagoko tarteak eratuko ditugo.

- Metodo hauek aztertu ondoren ikus dezakegu *Sturgesen metodoa* oso erabilgarria dela, *Limiten Teorema Zentralean* lantzen delako. Horrela, batazbestekoari gertu dauden tarteen dentsitatea haunditzen ditu, tarte kopuruak haunditu baino lehenago.

- *Scott* eta *Freedman Diaconisen* ereduak, oso sentikorrak izango dira datu atipikoak dauzkagunean, hauen kalkulurako desbideratzea eta kuartilen arteko ibiltartea erabili behar delako. Beraz, datu atipikoak dauzkaten grafiketan, komenigarria izango da, método hauek ez erabiltzea.

Irudia 2Tarteak definitzeko aukeren arteko gida Iturria: egileak egina



2.2.1.1 Tarte erregularretako histograma

Adibide gisa, R programarekin datu batzuk bilduko ditugu. Datu hauek, programak berak sortutako datu aleatorioak izango dira[10]. Suposatuko dugu, Diputazioak, errepide batean 80 km/hko radara jartzea aztertzen ari dela, uste duelako kotxeak azkarregi pasatzen direla eta hartutako abiadurak hauek izan direla.

Taula 2: sample(50:100,160,replace=T) komandoarekin, 50 eta 100 artean jasotzen diren datu aleatorioak sortuko ditugu. R programa erabili dugu, datuak sortzeko.

[1] 60 86 95 99 65 63 67 85 66 73 97 53 95 91 55 70 88 91
[19] 85 94 57 71 64 93 95 61 76 100 70 87 94 89 78 89 66 86
[37] 65 75 50 61 93 55 95 75 65 90 55 81 87 53 85 93 55 88
[55] 86 84 88 55 51 96 50 92 92 97 68 85 81 69 78 52 99 50
[73] 57 85 88 57 63 85 96 92 70 81 84 51 83 74 50 87 98 90
[91] 70 59 56 76 58 54 90 94 95 88 75 73 100 99 94 54 51 98
[109] 50 65 95 78 79 93 51 58 80 52 54 69 59 92 64 82 76 64
[127] 67 50 80 91 89 54 63 58 50 86 68 72 76 82 99 89 87 92
[145] 52 54 51 74 76 85 89 94 58 77 52 64 61 74 61 97

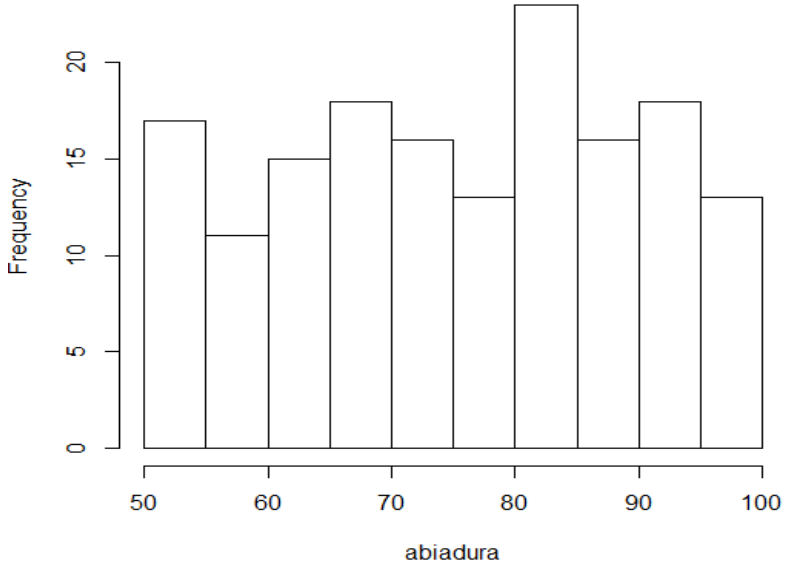
Sturgesen erregelaren arabera, 160 datu elkartzen badira, 9 tarte finkatu beharko ditugu. Tarte hauek erregularrak direla suposatuz, hau da, luzera berdina dutela. Beraz, lehenengo tartearen hasiera, datu baxuenetik hasiko da, eta bukaera, datu altuenarekin bukatuko da. Datu baxuena 50 km/h denez eta altuena 100 km/h denez, ibilbidea 50 km/hkoa izango da. 50 km/h hauek 9 tartetan biltzen baditugu, 5,55 km/hko tarteak eratu beharko dira. Nahiz eta 5,55km/hko tarteak eratzea gomendatu *Sturgesen erregelak*, histograma ulergarriagoa izateko, tarteak 5 km/htara borobilduko ditugu bi arrazoiengatik:

1. Tarteak ulergarriagoak izateko, komenigarria izango da tarteen luzeera borobiltzea. Kasu honetan, 5eko multiploak, azkarrago ulertuko ditugu, 6ko multiploak baino.

[10] `> sample(50:100,160,replace=T)` komandoarekin, 50 eta 100 artean jasotzen diren datu aleatorioak sortuko ditugu. R programa erabili dugu, datuak sortzeko.

2. Lehen aipatu dugunez, oso garrantzitsua izango da datuen aurkezpena zertarako izango den ulertzea. Kasu honetan, Diputazioak aztertu nahi duena, zenbat kotxe pasa duten 80km/h orduko abiadura izango da, eta horrekin batera ze abiadura tarteetan. Ez da berdina izango 82 km/h edo 98. Beraz, garrantzitsua izango da histogramaan, 80 km/htan, tarte bat bukatzea, horrela, azterketa zehatzagoa izan daiteke, nahiz eta tarte optimoen aipatutako erregela guztiz ez errespetatu. Beraz, histograma[11] horrela geratuko da:

Irudia 3Gure zentsutik lortutako abiadurak Rrekin sortuta. Iturria: Egilea



Histograma honetan ikusi daitekeenez. Jende askok finkatutako abiadura baino azkarrago pasatzen dela ikusi daiteke, beraz, radar bat jartzea komenigarria dela ondorioztatu daiteke.

Honako hau histograma arrunta bat izango da. Urteak igaro diren bezala, estatistika eta histogramak garatzen joan dira. Garapen honek histograma mota desberdinak ekarri ditu. Ikusitako kasu honetan, tarte erregularretako histograma aztertu dugu, baino zer gertatuko da histograma batekin, zeinek tarte irregularrak dituen? Oso garrantzitsua izango da, histograma hauek ulertzen jakitea, eta eratzen ikastera. Ikus dezagun, beraz, tarte irregularretako histograma bat nola eratzen den:

[11] Hist(abiadura) komandoa sartu beharko dugu R programan, ondoren ikusiko dugun bezala.

2.2.1.2 Tarte irregularretako histogramak

Datuak grafikora pasatzean, batzuetan, datu hauek alborapen handia izango dute. Honek, gure histograman hutsuneak sortuko ditu, eta hutsune hauek ez ikusteko, histograma batean, tarte irregularrak sortuko ditugu. Komenigarria izango da, ahal bada, emaitza hobeto ulertzeko, tarteak erregularrak izatea, baina kasu batzuetan, ezinbestekoa izango da era honetan egitea. Prozesua, histograma normal batena bezalakoa izango da; Datuak bildu ondoren, tarteak eta taula eratu, eta honekin grafikoa sortu. Histograma hau osatzeko, kontuan izan beharko dugu maiztasun erlatiboak erabili beharko ditugula altuerak eratzeko. Beraz, histograman tarte batek izango duen altuera, tarte honetan datuen kopurua guztiarekin zatituz eta hau tartearen luzerarekin zatituz lortuko dugu. Adibide batekin, noski, errazago ulertuko dugu:

Imajina dezagun datuak biltzen dituen taula batean, honelako datuak aurkitzen ditugula:

TARTEA	N
1-50	35
50-250	35
+250	70

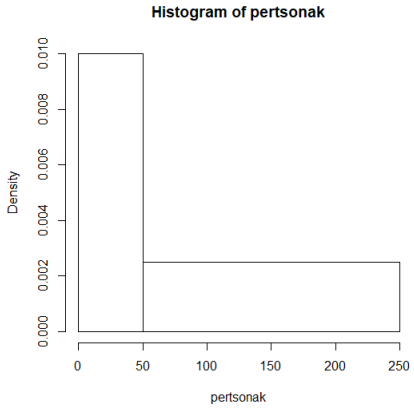
Taula 3Histograma irregularraren errepresentaziorako sortutako datuak. Iturria: egilea

Beraz, suposatu daiteke histograma bat eratzerakoan bi tarteak altuera berdina izango dutela, baina orain aipatu dugun bezela, datuak irudikatzeko formula aplikatu behar dugu. Modu honetan, datuen trinkotasuna ebazten dugu:

ZUTABE ALTUERA=MAIZTASUNA/ZABALERA

Lehen bezala, R programa erabili ezkero[12], honako grafika hau lortuko dugu:

Irudia 4Histograma irregularraren kasua Rrekin landua. Iturria: egilea



Histograma batetik, pentsatzen duguna baino informazio gehiago lortu dezakegu, denbora labur batean. Datu errepikatua zein den ikusteaz aparte, datu hauek jarraitzen duten distribuzioa ere ulertzeko erraztasuna emango digu, estatistikako tresna honek. Horretarako, argi izan beharko dugu histograma baten helburua, maiztasun haundiena duen elementua zehazteaz aparte, beste erabilera ezberdinak izango dituela. Ondoren, labur azalduko dugu identifikatu daitezkeen distribuzioak eta hauek zer esan nahi duten. Ikus dezagun, beraz, ze histograma mota eratu daitezkeen:

Histograma simetrikoa:

Izenak berak esaten duen bezala, kanpai forma izango duen banaketa jarraituko duen diagrama izango da. Kasu honetan, bataz bestekoa, moda eta mediana antzekoak izango dira eta datuak distribuzio normala izango dute. Beraz, azterketa estatistikoetarako, distribuzio normalarekin bat datozten formulak aplikagarriak izan daitezke.

Alboratutako histograma:

- Ezkerrera alboratutakoa: Grafiko hau eratuko da, aztertzen dugun aldagaia, balore bat baino haundiagoa izatea ezinezkoa denean. Adibidez, pila

[12] `>hist(pertsonak,breaks=c(0,50,250))` komandoa erabili dugu, datuak sartu ondoren

marka bakoitzak sortzen duen erradiazioa neurtzean, maiztasun gehiena erradiazio baxuena duen tartea izango da.

- Eskubira alboratutakoa: Kasu honetan, aldagaia balio bat baino baxuagoa ezin daitekeenean izan.

Histograma bimodala:

Kasu honetan, histogramak bi distribuzio normalen elkarketa bezala definitu dezakegu. Erdian, bailara bat bezelako fotma sortuko da, eta alboetan datuak oso gutxi errepikatuko dira. Adibidez, makina baten funtzionamendua eratzeko irudikatu daitekeen irudia izan daiteke.

Moztutako histograma:

Beherako edo gorako malda jarraitzen duen distribuzioa. Batzuetan, grafika mota hau horrela eratzen da, datuak biltzean errore bat egin delako.

Histograma laua:

Aztertutako datuek banaketa uniforme bat jarraitzen dutenean, maiztasunak tarte guztietan antzekoak izango dira. Beraz, histogramak ez du malda definiturik izango eta maiztasun guztiak antzekoak izango dira.

Muturreko histograma:

Histograma batek tendentzia edo banaketa bat jasotzen duenean dirudienez, eta muturreko tarte batean, bai ezker edo eskuinean, maiztasuna oso ugaria da.

Orrazi histograma:

Histograma batek forma hau jasotzen duenean, agian, histograma eratzeko jaso diren datuak, era okerrean biribildu edota tarteak finkatu direla esan daiteke.

Moztutako histograma:

Batzuetan, biltzen ditugun datuetatik, batzuk ez ditugu sartuko. Beraz, errepresentazio grafikoan, agertzen diren datuak baldintza bat betetzen dutena izango dira. Hau gerta daiteke, adibidez, kalitate kontrola pasatzen duten produktuen histograma bat eratzean. Naranjen pisua aztertzean adibidez.

Naranja oso txikiak baztertu egingo dira, salmenta prozesurako balio ez dutelako.- Era honetan, gure laginean hutsune bat izango dugu, 0tik hasita, merkatuak aprobatzen duen pisu minimoraino.

Histograma motak:

Dentsitate histogramak eta probabilitate banaketak:

Datu batek, histogramak dituen tarteen batean egoteko probabilitatea adieraziko digun grafikoa izango da dentsitate funtzioa. Grafiko honekin batera, datuen banaketa aztertu dezakegu.

Maiztasun poligonoak:

Maiztasun diagrama bat, histograma baten ondoren sortu daitekeen grafikoa da. Grafiko hau, histogramaren zutabeen tarteen erdiko puntuak marra batekin lotuz eratuko da. Histograma normalaz urruti, beste grafika batzuekin ere erabili daiteke grafiko eratorri hau. Adibidez, maiztasun erlatiboak aztertzean, datuak jarraitzen duten distribuzioa errazago ulertzea laguntzen digu.

Grafiko mota hauek, bi histograma alderatzeko oso erabilgarriak izan daitezke, errazago adierazten direlako bi maiztasun poligono batera, bi histograma batera baino.

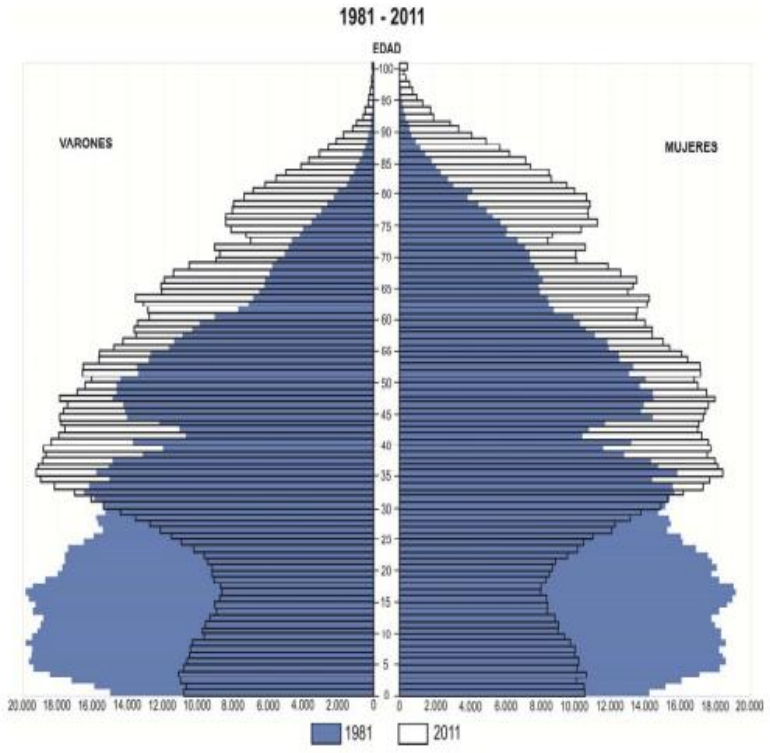
Baterako histogramak:

Estatistika deskribatzailearen helburuetako bat, datu guztiak grafikoki deskribatzea hauek beste datu batzuk sortzen dituzten grafikoekin konparatzea izango da. Beraz, datu ezberdinen histogramak sortu ezkero, eta hauek erlazioa izan ezkero, oso baliagarria izango da grafiko hauek bateratzea, aztertzailearentzat eta irakurlearentzat, datuak era laburrago batean ulertu ahal izateko.

Bateratutako grafikoen kasu asko ikus ditzakegu, baino histogramekin zerikusia duen bateratutako grafiko ezagunenetako bat, biztanleria piramidea izango da. Grafiko hau, era sinple batean egin daiteke, gizonak eta emakumeak bateratuz, edota bakoitzak bere grafikoa izanez. Hauek bateratu nahi ezkero,

noski, tarte berdinak erabili beharko ditugu. Aipatutako lehenengo grafikoan, (gizonak eta emakumeak grafika berean dauzkagunean) ondorio globalak atera ditzazkegu, eta bestean (gizonak alde batetik eta emakumeak bestetik) konklusio zehatzagoak aterako ditugu. Ikus dezagun, grafiko hauek bateratzen dituen biztanleria piramide bat. Datuak zehatzak izateko, gomendagarria izango da portal ofizial batetik ateratzea. Gure kasuan, Euskal Herriko biztanleria piramidea aztertu nahi dugunez, *Eustat*[13] estatistika portalean begiratuko dugu. Bertan, demografia aztertzen duen *dossier*[14] bat aurkitu dezakegu, 2018 urtekoa. Hemendik aterako dugu hurrengo irudia, non Euskal Autonomia Erkidegoko biztanleria piramidea ikus daitekeen, eta datuen irudikapena aztertuko dugu:

Irudia 5: Euskal Autonomia Erkidegoko biztanleria piramidea iturria: EUSTAT



[13] EUSTAT. (2020). *Datos estadísticos de la C.A. de Euskadi*. Eustat.eus. Iturria http://www.eustat.eus/indice.html. euskal Estatistika Erakundea-Instituto Vasco de Estadística bezala ezagutzen dugun institutu públiko ofiziala izango da. Gobernuak dauzkan datuak, biztanleak era sinpleago batean informazioa jasotzeko daukan erakundea izango da.

[14] EUSTAT. (2018). *PANORAMA DEMOGRAFICO 2018* (p. 10). Eustat.eus. Iturria https://www.eustat.eus/elementos/ele0015200/Panorama_demografico_2018/inf0015282_c.pdf Bertatik lortu dugu lan honetan 5. Irudia bezela identifikatuta dagoen biztanleria piramidea.

Aurreko biztanleria piramidean ikusi daitekeenez, gizonen eta emakumeen banaketa antzekoa izango da. Argi ikusten da, emakumeek adibidez, bizitza esperantza handiagoa izango dutela (piramidearen goikaldean emakume gehiago nabarmentzen dira, gizonak baino). Konklusio hau, adibidez, ez genuke lortuko gizon emakume desberdinketa erabiliko ez bagenu. Adibidez, lortu genezakeen ondorioetako bat, biztanleria zaharkitzen ari dela eta jaiotza tasa baxuegia dagoela izango zen. Honek pentsio sisteman izango duen kalte nabaria suposatuko duenarekin, adibidez. Edota pertsona baten bizitza esperantza zenbatekoa den sexua kontuan izan gabe. Esan bezala, grafikoak bakoitza alde batetik egin ezkero eta emaitzak bateratu ondoren, ondorio zehatzagoak ateratzea lortuko dugu. Aprobetxa dezagun grafiko hau bera, histogramen eraketan egiten diren akatsak berrikusteko:

- Lehen aipatu dugun bezala, komenigarria izango da, tarteak eratzeko sistemaren bat erabiltzea. Kasu honetan, urte bakoitzak tarte bat izango du, beraz, tarte gehiegi egongo dira.
- Beste grafiko bat antzeman daiteke, 1981 urteko biztanleria piramidea izango dena. Lehen aipatu bezala, maiztasun poligonoak konbinatzea errezagoa suertatzen da, bi histograma baino, hauek solapatu egiten direlako. Aztertzaileek akats hau egin dute, eta ez da oso ondo nabarmentzen 1981eko biztanleria piramidea eremu batzuetan.

Ikusi daitekeenez, oso garrantzitsua izango da histograma batean, ze histograma mota eta tarteak definitzea. Hau egin baino lehenago, horregatik gomendatzen da tarteak eratzeko sistemaren bat jarraitzea, lehen aztertu dugun bezala.

Ikusi dugun bezala, histogramak eratzeko ez dira eragiketa matematiko konplexuak burutu behar, baina oso adi egon behar gara hauen eraketan, grafikoak irudikatzeko aukera ezberdin asko izango baititugulako, eta aukera guzti hauek matematikoki era zuzenean egongo dira planteatuta. Beraz, irakurlearen azalean sartu beharko gara, eta datu guzti hauen azterketa zertarako izango den aztertu beharko dugu.

Eratutako grafiko hauek, noski, ordenagailuz programa ezberdinekin eratzea daukagu. *Excel* berarekin grafikoak eratzeko aukera izango ditugu, baino gauza konplexuagoak burutu nahi ezkero, programa informatiko espezializatuak bilatu beharko ditugu. Ondoren, programa hauetako bat ezagutuko dugu, R izenarekin ezagutzen dena. Programa hau, kode irekiko programa bat da. Honek esan nahi du, edonork proposatu ditzazkeela hobekuntzak eta internet bidez programan burutu dituzten hobekuntzak aurkezteko aukera izango dute. *Linux* programa informatikoaren sortzaileak[15] zioen bezela, *"begi askorekin, akats guztiak ekiditeko gai gara"(Raymond, E.(1999))*. Honekin, argi uzten da, zenbat eta programa garatzeko erabiltzaile gehiago izan, hobekuntza nabarmenagoak garatzeko aukerak izango ditugula. Beraz, garrantzitsua izango da, eguneratu egiten den soporte batekin lan egitea, eta akatsak ekiditeko aukera eskainiko badizkigu. Programa ezberdin asko dauzkagu merkatuan, funtzio hau bete dezaketenak. Hurrengo atalean, programa ezberdin hauek aztertuko ditugu, abantailak eta desabantailak aztertuz. Guk aukeratu dugun programa, R *softwarea* izan da. Kosturik gabeko *softwarea*, kode irekikoa,

Ikus dezagun beraz hurrengo atalean, R programak historikoki izan duen garapena, eta zergatik aukeratu dugu programa hau gure lana garatzeko, beste programa batzuekin konparaketa ezberdinak eginez.

Irudia 6 Rren logoa.
Iturria RPubs

[15] Raymond, E.S.. (1999) *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, Cambridge (MA), O'Reilly. Rasymondek esan nahi duena, soluzioa daukaten akats asko, edonorkek erreparatzeko aukera izanda, programaren funtzionamendua efizienteagoa izango dela.

3 R Project programa informatikoa

3.1 Garapen historikoa

Teknologia denboran zehar garatzen joan den heinean, matematikak eta enpresek dituzten errekurtso informatikoak ere garatu dira. Hauen artean, programa informatikoek garrantzi haundia hartu dute empresa baten funtzionamenduaren barruan. Gaur egun, edozein enpresak erabili beharra dauka korreoa, *Microsoft Office* paketea, eta beste hainbat programa bere garapena zuzena izan dezan. Tresna informatiko ugari erabili ditzakegu, bakoitzak erabilpen ezberdina izango duena. Enpresa munduan, asko erabiltzen den programa R programa informatikoa izango da. Programa honek, grafikoak sortzeko aukera infinituak proposatzen dizkigu, eta grafiko hauek eraldatzeko aukera ditugu. Era honetan, empresa batek datu kopuru itzela hartu eta tratatzeko aukera izango du, edozein momentuan. R programa gehiago ezagutzeko eta nondik eratorria dagoen jakiteko[16], atzera begiratu behar dugu, XX mendean sortutako programa baita.

R programa Zelanda Berrian sortu zen, 1992 urtean *Ross Ihaka* eta *Robert Gentelmanengatik (Ihaka,1998)*. Programatzaile hauek zuten lehen helburua honako hau zen; hizkuntza dialektiko bat sortzea, zein estatistikako hastapenerako ikastaro baterako erabilgarria izan zitekeena Zelanda Berriko Unibertsitatearentzat. Horretarako, S programatzeko lenguaia eratorritako sintaxia erabiltzea erabaki zuten, *Bell Laboratoriesek* eratutakoa. Horregatik, programazio ezagutza haundia dutenek, S eta R hizkuntzak, anai-arrebak direla diote, bi hizkuntza hauen antzarengatik.

R izena, broma bat bezala hasi zen. *Ross* eta *Robert*-ek, laburdura hau erabiltzen zuten hizkuntza programatzaile hau definitzeko, beraien izenengatik, eta hemendik Aurrera horrela ezagutuko da hizkuntza. Hizkuntza programatzaile hauen garrantzia ulertzeko, S-ren garapen historikoari begirada azkar bat botako diogu. 2004 urtean, 2 milioi eurogatik saldu zen, eta 2008 urtean, 25 milioi euro ordaindu ziren, programazio lenguaia honen komertzializazio eskubideak lortzeko.

[16] https://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf

Itzul gaitezen R programara. R hizkuntza sortu ondoren, publikoki R *softwarea* 1993 urtean aterako da. Bi urte ondoren, 1995 urtean, eskola *Zurich*eko eskola politekniko federaleko *Martin Mächlerek*, bi programatzaileei *GNU* lizentzia lortzea gomendatuko die, *software* libre bat bihurtzeko.

"Erabiltzaileek, entorno interaktibo batean lan egiteko aukera bilatzen dugu. Ez ikustea programatzaile hutsak bezela, garapenean lan egiten duten pertsonak baizik. Zenbat eta behar haundiagoak izan, programazio ezagutza behar haundiagoak izan beharko dituzte. Bertan, kodea eta sistema garrantzitsuak bihurtzen dira" (Chambers eta Hastie [1991] egileak itzulita)

1997 urtetik aurrera, R *softwarea GNU* proiektua barne egongo da. Horretarako, urte bat lehenago, R programaren sortzaileak, beraien korreoa publikoki aurkeztuko dute, hizkuntzarentzat soporte baten sorkuntza lortzeko. Edonork bidali zezakeen bere bertsioa hizkuntza hau sortu zutenei, eta jendeak positiboki hartu zuen. Korreo gehiegi jasotzen zituztenez, korreo berriak sortzeko obligazioa izan zuten. Honen ondorioz sortzen dira horrengo bi korreoak, gaur egun oraindik funtzionamenduan jarraitzen dutenak; *R-help* eta *R-devel*. Korreo hauek, gaur egun, erabiltzaileen problemak konpontzeko aukerak proposatzen dituzte. 2000 urtean, ateratzen da R-ren bertsioa merkatura. R 1.0 izanarekin ezagutuko da.

Programak urte hauetan garapen ezberdinak izan ditu, eta erabiltzaileak urtero gehiago dira. Momentu oro garatzen den programa bat, eta doan dena, erabiltzaileek gustura erabiltzen dute. Baino, erabiltzaile arruntentzat bakarrik ez den programa bat izango da, baita ere, enpresek erabiltzen duten softwarea izango da, eta ez enpresa txikiak. Mundu mailan sektore ezberdinetan potentzia haundiak diren enpresek, *software* hau erabili edo garatu egingo dute, ondoren ikusiko dugun bezala.

R programaz aparte, beste programa informatiko ugari erabili ditzazkegu gure datuak bateratzeko eta grafikoak sortzeko. Badaude *software* batzuk, errazagoak izango direnak eta aukera gutxiago eskainiko digutenak,

Microsoft Office Excel edo *Libre Office Calc* programak adibidez, datu batzuk aukeratu ezkero grafikoak eratzeko aukera ezberdinak eskainiko dizkigu. Noski, ez digu R-k aurkezten dizkigun aukerak aurkeztuko, *Excel softwarea* beste erabilpen batzuetarako espezializatuta dagoelako. Ondoren, R programaren abantailak eta desabantailak aipatuko ditugu, eta bere konpetentzia diren programekin alderatuko dugu.

3.2 Nork erabiltzen du R?

2017ko irailean, *TIOBE programming community index*-ean (2017), mundu mailan dagoen indize ospetsuenetariko bat popularitateari dagokionez programazioaren munduan, R 11. lenguai erabiliena bezala finkatu zen. Urte bat lehenago, 18. postuan zegoen. Agian, ez da harrigarria hizkuntza bat mundo mailan 11. izatea, baina kontuan izan behar dugu honako hau: R programaren eremua, estatistika izango da, eta adibidez, beste hizkuntza batzuk (*Java*ren kasua bezala), edozein lanerako erabilgarri suertatu daitezkeen hizkuntzak izango dira. Beraz, Rk indize honetan igotzeko, muga handiagoak izango ditu, beste programa batzuk baino. Horregatik nabarmentzen da, hizkuntza hau munduko 11. erabiliena izatea. Urte honetan, 18. postura jeitsi da berriz ere, 2018 urtean 8. postuan egon ondoren. Oinean ikusi daitekeen web orrian[17], R hizkuntzaren garapen historikoa ikusi daiteke, indize honen arabera.

Rren igoera, batez ere, hizkuntza eta kode irekiaren erabilpenarengatik da. Honek, datuen erabilpen efektiboago bat erakarriko du, eta datuak prozesatzeko erraztasun haundiagoak eskainiko dizkigu. Gainera, lehenago aipatu dugun bezala, edozeinek ekarpenak garatzeko aukera izateak, eboluzio nabarmenagoa izatea lagunduko du. Kontuan izan behar dugu, nahiz eta dirua izan, empresa askok R programa erabiltzen dutela, beren lanak burutzeko. Adibide gisa, *Santander bankua* izan daiteke[18], banku honek bere langileen

[17] TIOBE Programming Community Index. (2017). *R | TIOBE - The Software Quality Company*. Tiobe.com Iturria https://www.tiobe.com/tiobe-index/r/. web orria, non Rren indizearen garapen historikoa ikusi daitekeen.

[18] COM UPM, & Santander. (2015). *Taller de aplicaciones en R*. Youtube. Iturria: https://www.youtube.com/watch?v=-m6XcgfUKRQ. Link honetan, Santander bankuan parte hartzen duten langileek, R buruzko ikastaro bat jasotzen dute. Ikastaro hau youtuben ikusi daiteke, doan.

formazioa bilatzen du, eta arlo batzuetarako beharrezkoa duten R hizkuntza irakasten dute. Beste kasu nabarmen bat, *Facebook*en kasua izango da. *Facebook*ek, erabiltzaileek beren orrialdeetan egiten dituzten interakzioak ikertzeko erabiltzen dute programa, era honetan, zer nolako publikazioak erakutsi ahal izateko jakingo dute. Kontuan izan behar dugu, *Facebook*en diru sarrera garrantzitsuena hau izango dela, erabiltzaileek publizitatearekin izango dituzten interakzioak. Horretaz aparte, *Faceebook*en giza baliabideen arloak ere erabiliko du programa hau, bere langileen interakzioak ikertzeko. R programarekin *facebook*eko edonorken interakzioak ikertzen irakatsiko digu, Oinean ikusi daitekeen linkean[19]. Nahiz eta histogramekin zerikusi haundirik ez izan, oso aipagarria da *marketing* arloan izan dezakeen ekarpena.

Google-ek ere erabiliko du, bere kanpainen efizientzia neurtzeko, bere eskainitako zerbitzuetan. Adibidez, ordaindutako publizitatea guk zerbait bilatzean edo *googleatzean*. *Facebook*en kasuarekin batera, bere irabazien zati garrantzitsu bat, modu honetan datozte. Gainera predikzio ekonomikoak egiteko erabiliko dute.

*Microsoft*en kasua desberdina da. Enpresa honek, bere bertsio propioa garatu zuen, *OpenR* izenekoa. Egunean, erabilpen irekikoa dena. Bertsio hau, analisi estatistiko ezberdinak garatzeko erabilia da. Adibidez, *XBOX* erabiltzaileak bere mailako jokalariekin enparejatzeko erabiltzen du, R hizkuntza, beste kasu ezberdinen artean.

Esan dugun bezala, enpresa askok erabiltzen dute R egunero, ez da errekurtsorik gabeko jendearentzat bideratutako programa bat. Kasuak aipatzen jarraitu dezakegu, *IBM, Ford, American Express, ...* aipatu ditugun kasuak, Rren aplikazio espezifiko ezberdin batzuk irudikatzeko baliabide gisa erabili ditugu. Esan dezakegu beraz, Rk datuak, prozesatzeko, ikertzeko, moldeatzeko eta komunikatzeko balio duela. Bukatzeko, mundu mailan erabiltzen duten enpresen lista bat utziko dugu

[19] Barberá, P. (2017). *Access to Facebook API via R*. Rfacebook paketearen gidaliburua: https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf Rfacebook paketeak eskaintzen dituen komando guztien azalpena.

"Hurrengo konpainiak R sistema erabiltzen dute, jarduera hauetan:

1. Facebook, perfileko argazkien informazioa biltzeko eta aztertzeko.
2. Google, anuntzioen efizientzia aztertzeko
3. Twiter, tuit guztien erreperkusioa aztertzeko
4. Microsoft barne garapenerako lanak burutzeko
5. Uber analisi estatistikoak burutzeko
6. Airbnb prezioak aztertzeko
7. IBM Rrekin produktuak sortzeko
8. ANZ kreditu arriskuak aztertzeko
9. HP
10. Ford
11. Novartis
12. Roche
13. New York Times datu bisualizaziorako
14. Mckinsey
15. BCG
16. Bain

Nahiz eta R, printzipioz, erabilpen estatistikorako sortua izan, erabiltzaileak programa honen potentziala esplotatzeko gai eta errendimentu haundiagoa lortzeko gai izan dira. Programa edonork garatzeko aukerak, pakete ezberdinak garatzeko aukera erraztu du, eta horregatik, gaur egun, R edozein aplikaziorako erabili daiteke. Datuak gordetzeko eta editatzeko, irudien prozesamendurako, irudi interaktiboak sortzeko, *Big Data* prozesamenduan, eta beste aspektu ezberdinetan.

- R kalitatezko *software* bat da. Oso diseinu ona daukan programazio lenguaia bat da, bai perspektiba analitikotik (estatistika) eta konputazionaletik (informatika). *Fox* eta *Andersen* adituek (2005) esaten duten bezala. Baieztapen hau, beste batzuen artean, abalatuta dago, 1998 urtean, John Chambersek jasotako sariarekin *Association for Computing Machinery* – Informatikoen asoziazio nabarmenena – S hizkuntzaren garapenarengatik (Rren erabilpenerako beharrezkoa den hizkuntza). *Chambers*, gaur egun Rren garapenean lan egiten duen aditua da *(R core*):

ikerkuntza estatistikoaren beste figura nabarmen batzuk R core taldearen barnean aurkitzen dira. Adibidez, *Douglas Bates, Brian Ripley* edota *Luke Tierney*.

- R doaneko programa informatikoa da. Gure garunean, automatikoki sortzen den erlazioa honako hau dela baieztatu daiteke: benetan ona izango da? Aditu askok psikologikoki gizakiak garatutako sentipen honi buruz ikerketak egin dituzte, eta oso arrunta da sentimendu hau izatea. Zorionez, informatika munduan asoziazio hau ahuldu egin da, kodigo irekien iniziatiba ezberdinak burutzerakoan. R, kode irekiko aparteko programen multzoaren barruan dagoela esan daiteke, eta ez dauka ezer enbidiatzeko bere merkatuko beste produktu batzuekin, hauek kostu altukoak izanda. Gainera, Rk daukan abantailetako bat, bere erraztasunak izango dira. Beste programen dokumentazioa bilatzea, Rn baino kostu haundiagoa izango du, eta R plataformak manual espezifikoak izango ditu, hizkuntza ezberdinetan. Gaztelera, ingelesa, frantsesa…Vietnamerera ere irakurri daitezke Rrako prest dauden manual ofizial ezberdinak. Beraz, esan daiteke R mundu mailan zabaldu den programa informatikoa dela, hizkuntza barrerarik ez duena eta prezio mugarik ere ez. Edonork erabili dezakeela.
- R *softwarea*, multiplataforma da. Honek esan nahi du, edozein ordenagailutan erabiltzeko aukera izango dugula. Orain dela urte batzuk, *Window*seko erabiltzaileeek ezin zituzten *MacO*seko programak erabili, eta alderantziz. Nahiz eta problema hau, oso nabaria ez izan orain egun, kasu batzuk mantendu egiten dira, bereziki *LINUX softwarea*ren erabiltzaileek izaten dituzte. Hauek, beste programa batzuen antzeko programak erabili behar izango dituzte, lan ezberdinak burutzeko. Rk restrikzioak dauzka, kasu hauek ekiditeko eta sistema operatibo ezberdinetan efizientzia berdina izateko. Modu honetan, edozein erabiltzaierentzat prest egongo den programa bat izango da, hau da, guztiz eskuragarri eta erabilgarri den programa informatiko bat izango da.
- Rk programatzeko aukera emango digu. Are gehiago, Rk programazio hizkuntza bat garatzen du, eta ez bakarrik komando bateratu bat. Honek esan nahi du, programa hau ez dela mugatua izango, baizik eta ia edozein

gauza burutzeko aukera emango digu. Hau da, potentzialki, ez dugu limitaziorik izango, operazioak burutzeko. Adibide bezela, *Ruiz Soler-ek dioenez* (2005).

"Maiz eskatzen dira lan edo prozedura ugari, zeintzuk nahiz eta ez izan prozesu estatistiko garbiak, beharrezkoak dira ikerketa lanetan; kasua da adibidez, kontrabalantze sekuentzien sorkuntza, diseinu esperimentalen barne aldagaien erroreen progresioa neutralizatzeko (nekea, ikasketa edo influentzia hondakinen eraginak). R programaren erabilera erraz automatizatu ditzazke horrelako prozedurak, zeintzuk bestera, lan karga astun bat lirateke aldagai eta baldintza ugari parte hartzen dutenean. (Ruiz Soler eta Lopez Gonzale (2009[20]) egileak itzulita).

Hau da, R programak, erroreak sortu dezaketen datuen neutralizazioan laguntzeko aukera mugagabeak emango dizkigu.

- R denborarekin hazi egiten da. Rren garapena itzela da. Bere lehenengo agerpenetik, 2000 urtearen otsailak 29an, bertsio desberdin berriak sortu dira, hileroko periodizitatearekin. *Ripley* adituak esaten duenez *(Ripley, 2003, Ruiz Soler eta López González-en testutik errekuperatuta):*

  "Izan daiteke, programa honek eskaintzen dituen aukera guzti hauek, gutxi batzuen beharrak asetzeko sortuak izatea. Era honetan, Komunitate guztiari iritsiko zailo garapena. Zenbat eta denbora aurrera doan, zenbat eta materia gehiagotan ikusi dezakegu R programa aplikatzen dela. Mugarik gabeko programa dela. (Burns (2007), Ruiz Soler eta López González-en testutik errekuperatuta, egileak itzulita).

  Momentuan, Rk 400 pakete baino gehiago dauzka eskuragarri. Hauek Area ezberdinetarako prestatuta dauden paketeak dira. Area ezberdinetarako prest dauden datu baseak eta eragiketa espezifikoak burutzeko komandoak barneratuko dituzten pakete hauek, Rren garapenaren elementu garrantzitsuak izan dira.
- Rren grafikoen kalitatea bikaina da. Adituen ustez, hau izan daiteke Rren erabilpena sustatu duen elementu nabarmenetariko bat. Ez dute

[20] Ruiz Soler, M. eta López González, E. (2009). *El entorno estadístico R: ventajas de su uso en la docencia y la investigación*. Granada. Granadako unibertsitatea. Bertan Rren abantailak aztertzen dituzte, eta honen irakaskuntza gomendatzen dute.

konpetentziarik beste antzeko programekin (baieztapen hau kontsultatu nahi ezkero, web orri hau kontsultatu dezakete: http://addictedtor.free.fr/graphiques/, non ehundaka grafiko ikus daitezkeen. Adituen esanetan:

"Uste dugu grafikoak bereziki garrantzitsuak direla grafiko kategorikoentzat, nahiz eta oraindik ez diren asko zabaldu. Bere abantailak datu interpretaziorako kontingentzia tauletatik eratorrita izugarriak izango dira. (Friendly eta Ruiz Soler (2000), Ruiz Soler eta Perez Garcia, egileak itzulita)

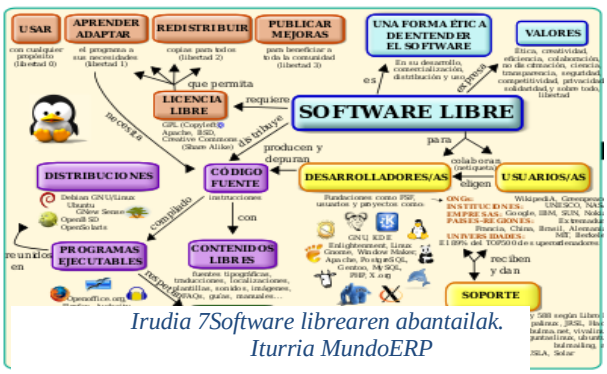
- Rk matematikoki pentsatzen laguntzen du. Komandoak erabiltzeko pentsatutako interfazak, ikasle edo erabiltzaileei zer egiten hari diren pentsatzean behartuko die *(Ripley, 2003)* eta autore ezberdinek erakutsi dute estatistika ikastaroetan nola erabili daitekeen R (adibidez, *Nolan* eta *Speed*, 2000); baita ere ikus ditzazkegu lan ezberdinak, non R, kontzeptu estatistikoen esploraziorako lan tresna bezala erabili daiteke.

Abantaila hauek, *Marcos Ruiz Soler* eta *Evelina Gonzalez Lopez*ek irudikatu zituzten, Rri buruz egindako ikerketa batean. *El entorno estadistico en R: Ventajas de su uso y Docencia en la investigación (2009)*, mundu akademikoan izan ditzazkeen abantailak irudikatuko dizkigu. Honetaz aparte, Rrekin egin daitezkeen eragiketa edo operazio matematiko ezberdin batzuk ikusi daitezke. Bibliografian eskuragarri utziku dugu dokumentua, edozein irakurlek, nahi ezkero, informazio gehiago izan dezan.

3.3 R *softwareak* dauzkan abantaila eta desabantailak

3.3.1 R *softwareak* eskainitako abantailak

R programa, lan honetan askotan aipatu den bezala, sistema libre bat da. Honek esan nahi du *Software* libre batek eskaintzen digun abantaila, bere kodigoa aldatu daitekeela. Beraz, sistemak berak jartzen dizkigun oztopoak ebaztu daitezke. R



Irudia 7Software librearen abantailak. Iturria MundoERP

programak, 4 askatasun eskainiko dizkigu. Hurrengo irudian ulertu dezakegu zelako ondorio dauzka, *software* libre batekin lan egiteak:

1. Edozein moduan eta edozein helbururekin lan egiteko gai izango gara.
2. Programaren funtzionamendua ikusteko aukera eta gure beharretarako prestatzera.
3. Beste erabiltzaileek gure modeloak erabili ahal izateko kopiak gordetzeko aukerak eskaintzen dizkigu.
4. Programa maneiatzeko, eta komunitatearekin partekatzeko aukerak bere web orrialdean.

- R *softwarea*, beste *software* batzuekin konparatu daiteke. Desberdintasuna, prezioan nabaritzen da baita ere. Baliabide gutxi dauzkan enpresa batek, askotan, kostu baxuko programak beharko ditu. Hurrengo taulan, grafikatzeko programa ezagun batzuk ikusiko ditugu eta prezioen konparaketa burutuko dugu:

Taula 4 Programa ezberdinen prezioak. Iturria: egilea

SOFTWAREAREN IZENA	PREZIOA
R	0€
MATLAB	2100€
MINI TAB	1500€
SIGMPLOT	900€

- Plataforma ezberdinetan erabilgarri izango da. Honek esan nahi du, edozein erabiltzailek, edozein ordenagailutan erabili ahal izango duela. Batzuetan, programa askok duten oztopoa, sistema operatibo bakar baterako prest daudela izango da. Era honetan, edonork hartu ditzazke guk burututako lanak, bere garapena jarraitzeko edo ulertzeko.

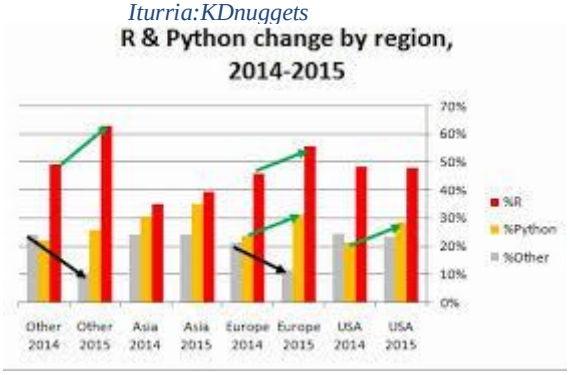
- R komunitatea 6000 erabiltzaile baino gehiagoz osatuta dago. Erabiltzaile hauek, programan burutzen dituzten hobekuntzak, eskura uzteko aukera dute. Adibide bezala, imagina dezagun gure programa garatu dugula, enpresa munduan erabiltzen diren datu estatistikoen tratamendu espezifikoa garatzeko. Interneten, gure lana utzi dezakegu, beste erabiltzaile batek, guk egin dugun lana aprobetxatu dezan bere onurarako

- R *softwarearen* eboluzioan orduro lan egiten duten programatzaileak daude. Aktualizazioak askotan ikusiko ditugu, programa era optimoan funtziona dezan. Azkeneko bertsioa,4.0.2[21] bertsioa izango da.

- Lehen aipatutako abantaila bat, prezioa izan da. Bertan, Rren kostua beste *software* batzuekin konparatu dugu, baino kontuan izan behar dugu, R programak, beste funtzio batzuk osatzen dituzten beste programa batzuen lanak ere burutzeko gai izango dela, beraz, informazio guztia bateratzeko erraztasun haundiagoa izango dugu, eta dirua eta denbora aurreztu ahal izango dugu programa informatiko ezberdinak lortzen.

Rren garrantzia sozietate zientifikoarentzat handitzen ari da urtetik urtera. Zientzialariek, hauen lanetan erabilitako grafikak R programa erabiltzen dute. Programa honek, izen ona hartu du komunitate honetan, lanen estandarizazioaren bilakaeran, R programa erabiltzea gomendatzen baita, metodo estatistikoak erabiltzean. Hurrengo irudian antzeman dezakegu bere garrantziaren garapena, 2014tik 2015 urtera, kontinente ezberdinetan banatuta.

Irudia 8 R eta bere konpetentziak izan duten garapena azkeneko urtetan. Iturria:KDnuggets



Ikusi dugun bezala, R programak eskaintzen dizkigun abantailak ugari dira. Enpresa txiki batek erabili dezake kostuagatik, ikertzaileak erabili dezakete hauen lanak publikatzeko, enpresek erabili dezakete,

[21] 2020ko ekainaren 24ean eguneratutako informazioa. Bertsio berria dagoen ikusteko, hurrengo linkera bideratu. https://cran.r-project.org/bin/windows/base/

datuak partekatzeko,…eta hau guztia ezer ordaindu gabe aurkitu dezakegu.Gainera, *Europa* mailan erabilgarritasun haundia duen programa batean bilakatzen ari da. Goiko irudian ikusi daitekeen bezala, beste *softwareen* erabilpenaren gainetik kokatuta dago, *software* printzipala bilakatuz bai *Europa, Asia* eta *Estatu Batuetan*. Ikus dezagun R programan aurkitu daitezkeen oztopo edo desabantailak.

3.3.2 R softwarearen desabantailak

- R programak, gida ezberdinak eskaintzen ditu, bere web orrialdean. Dauzkan programa edo pakete bakoitzerako gida, oso zehatza izango da. Honek esan nahi du, pakete edo Rren funtzio baten informazioa bilatzeko, ez dela erraza izango informazio hau aurkitzea. Erabiltzaileak, ez ditu erraztasunak izango, programarekin laguntza jasotzeko.

- Bestetik, komandoa gaizki sartu ezkero sortzen diren mezuak, ez digu espezifikoki esango, zer egin dugun gaizki. Agertzen zaigun mezua, programazioari buruz gutxi jakin ezkero, ez da ulergarria izango. Honek esan nahi du, programazioriari buruz ezer ez dakienak, problemak izango dituela, komandoren bat gaizki exekutatu ezkero.

- R *softwarea*, komandoekin funtzionatzen du. Honek esan nahi du ez dugula menu bat izango. Berriz aurkitzen dugu programazioaren oztopoa. Beste programa batzuk, hauen dauzkaten menuko aukerekin, komandoak sartu gabe ekintza batzuk egiteko aukera eskaintzen digu. Horregatik, beste programa batzuetara ohituta dauden erabiltzaileei R programa erabiltzea ez dute gustuko izango.

Beraz, bere abantailak eta desabantailak ikusi ondoren, bere konpetentziarekin konparatuko dugu, eta programa garrantzitsuenekin alderatuko dugu.

3.4 R softwarea MatLabekin konparatuta

R programa, lehen aipatu dugun bezala, R hizkuntzarekin erabiltzen den softwarea izango da. Logikoa da, gu enpresa edo ikertzaileak izan ezkero, gure datuak partekatu ahal izatea, guztiek ulertu ahal dezaketen hizkuntza batean. Datuen analisirako software ugari aurkitu ditzakegu. Batzuk, beste programek izango ez dituzten abantailak izango dituzte, eta beste batzuk,

espezializazio bat izango dute. Programa guztien artean, *Matlab* programa aukeratu dugu, R softwarearekin alderatzeko. Kontuan izan behar dugu, *MatLab* programa, munduan dauden 25 unibertsitate prestigiosoenetatik 24tan erabiltzen dela[22].

Aipagarria den desberdintasun nagusia, eta erabiltzaile guztiek lehenengo begiratzen dutena prezioa izango da. *MatLab*en lizentziak 2100€ balioko ditu. Kostu hau oso nabaria izan daiteke empresa txiki edo ertain batentzat, beraz, enpresariak antzeko funtzioak burutzen dituzten programa merkeagoak bilatzeko joera izango du.

2017 urtean, Valparaisoko unibertsitateko ikasleek, *dossier* bat prestatu zuten, *R, Phyton* eta *Matlab* programak konparatzen zituena[23]. Ondoren aipatuko ditugun abantailak besteen aurrean ikertzaile hauetatik eratorrita daude, eta informazio gehiago orrialdearen oinean aurkitu dezakegun dokumentuan ikusi ahal izango da:

- *MatLab*en abantailak Rrekiko:
  - *MatLab softwarea*ren erabilpena, munduko unibertsitate prestigiosoenetan zabalduta dago (lehen aipatu dugun bezala). Gainera, aurrekontu haundiak dituzten enpresek, normalean *MatLab* programa hartzeko joera izango dute. *MatLab* programa, hasi berri diren erabiltzaileentzat errazagoa izango da. Ikertzaile hauek diotenez, programa hau erostean, behar dituzun pakete guztiak eskuragarri izango ditugu berehala, inongo komandorik idatzi behar izan gabe. *MatLaben* paketearen alderdi garrantzitsu bat, *Simulink* paketea izango da. *Software* espezializatuko pakete hau, bakarrik *MatLab* programan aurkituko dugu. *R* eta *Phyton* programetan ez dauden aukerak eskainiko dizkigu programa honek.

[22] Mathworks akademia. (2017). Es.mathworks.com. Iturria: https://es.mathworks.com/content/dam/mathworks/tagteam/Objects/9/90593_92300v01_TAHFactSheet_ES.pdf.
[23] Reed, D. (2014). *Data smart: Using data science to transform information into insight*. Journal Of Direct, Data And Digital Marketing Practice, 15. https://doi.org/10.1057/dddmp.2014.33

- Rren abantailak *MatLab*ekiko:
    - R, lehen esan dugun bezala, estatistikako programa bat da, eta estatistikako problemetarako lagungarria izango da. R *softwarea*, pakete askorekin dator, pakete hauek guztiak, gure problemak konpontzeko baliagarriak izango direnak. Hau izango da Rren funtzioa. *MatLab*ena berriz, matematikako beste aspektu batzuk, kalkulua eta ekuazioak bezala lantzea izango da ere. Beraz esan daiteke R programa, espezializatuagoa izango dela. *"R is currently widely used for data mining today" (Journal of data science 15 (2017))* ikertzaileek esaten dutenez, R programaren garapena, *software*aren erabilpenaren zabalpena erakarri du. Gaur egun, ikertzaileek esaten duten bezala, lan merkatuan kuota haundia hartzen ari da R programarekin lan egiten jakitea.

    "2017ko aktualizazio batean, Munichen, R programa datu analisiaren merkatuan izaten ari den gorakada nabarmentzen du. Enpresek eskaintzen dituzten lanetan, geroz eta aukera gehiagotan eskatzen dira programa honen ezagutzak izatea, datu analisiekin lan egiteko (R Bloggers, 2017)"

Arrazoi hauengatik, aukeratu dugu R programarekin lan egitea. Laburbilduz, programa merkea da, ikasle batek beharko duena, eta eskaintzen dizkigun zerbitzuak beste programen mailakoak izango dira. Gainera, honi batu behar diogu, lan merkatuan oso garrantzitsua izango dela. Honi batu behar diogu pakete espezializatuak izango ditugula, eta kode irekiko sistema izango dela *GNU* lizentziarekin. Ondoren, R softwareak izango dituen komandoen gida bat ikusiko dugu, datu sinpleak lortzetik grafiko landuak eratzeraino. Baino hau ikusi baino lehen, Rk eskaintzen dizkigun tresnak aztertuko ditugu, gure lana erraz dezan.

3.5 R Commander eta R Studio

R Commander, R barnean erabili daitekeen pakete bat izango da. Pakete honek, ez dizkigu funtzio berriak integratuko, baizik eta menú bat eskainiko digu, non komandoak idaztea baino errazagoa izango zaigun, *software*ari orden jakin batzuk ematea. Erabilera baxua duten erabiltzaileentzat oso gomendagarria den programa izango da. Noski, bere alderdi onak eta txarrak izango ditu.

Alde batetik, alderdi onen artean, denbora asko aurreztuko dugula izango da. Ez jakintzak programazioan denbora galera izugarri bat dakar, bereziki, komandoak idaztearekin zerikusia duten programekin. Menu bat izateak adibidez bataz bestekoa lortzeko, oso garrantzitsua izango da informazioa era labur batean lortzeko. Gainera, ez da inongo programa extrarik erabili beharko. Bakarrik, paketeen deskargen menutik, *RCMDR* paketea deskargatu beharko dugu eta ondoren liburutegitik inbokatu *library(Rcmdr)* komandoa zehaztuz.

Alderdi txarrak ere baditu. Adibidez, histograma eratzeko eskatzean, errazagoa izango da komando bidez zehaztea zein datu nahi ditugun grafikatzea. Noski, horretarako, programa honen ezagutza beharrezkoa izango da. Gainera, menú hau ez da denbora guztian hor geratuko, behin programa itxi egiten dugula, menú barra desagertu egingo da eta *library* komandoa erabili beharko dugu, aipatutako menua lortzeko. Gainera, menú honetan automatikoki sortzen diren komandoak, ez dira zehatzak izango. Honek esan nahi du, kasu batzuetan *R Comander*rekin lortu ditudan ekintzak, bere komandoak manualki kopiatu ondoren Rren kontsolan, ez ditudala emaitza berdinak lortu. Horregatik, komenigarria da programari erabilpen haundia emango bazaio, kontsolan programatzen jakiten ikastea eta kontsola ahalik eta gehien erabiltzea.

R Studio programaren kasua ezberdina da. Ez da pakete batean deskargatzen, beste *software* bat izango da, zein internetetik deskargatzeko aukera izango duguna. Programa hau, doakoa izango da sistema irekia bezela lan egiten baitu. Erabiltzaile askok programa honi buruz idazten dituzte, eta

askok *R Studio R Comander* baino hobeagoa dela esaten dute. Guztiak diote, erabiltzaile basikoenek lan egin ahal egiteko, *R Comander* paketearekin nahikoa izango dutela, baino %95eko lan egiteko ahalmen murrizketa izatea iritsi daitekeela. *R Studio*k paketea baino aukera gehiago ematen dizkigu, erabiltzaile berriei erraztasunak eskeiniz. Programa hau, erabiltzaileek oso programa konpletoa dela diote, baino gomendatzen dute azterketa sakonak egin ezkero, kontsolan komandoak manualki idaztea askoz aproposagoa izango dela beti, aukera nabarmen gehiago izango ditugulako. Programa honekin lan egin ondoren, esan dezakegu oso baliagarria izango dela, datuak kargatzeko eta datuen artean filtroak aplikatzeko. Era honetan, datu hauekin era egokiago batean lan egiteko aukera izango dugu.

Beraz, nahiz eta erabiltzaile berriak izan, ezinbestekoa izango da gure azterketetan informazioa lantzeko ahal diren aukera gehienak izatea. Horregatik, R *softwarea*rekin lan egingo dugu. Horretarako, gida txiki bat prestatuko dugu, non histograma bat grafikatzeko pausoak definituko ditugun eta honen edizioan aipagarriak izan daitezkeen paketeak aipatuko ditugu. Honekin sartu baino lehenago, programatzeko gomendio pare bat erraztuko ditugu. Programarekin lan egin nahi duenarentzat esperientzia errazagoa izateko.

3.6 PROGRAMATZEN HASTEKO AHOLKUAK

Programa bat hartzean, berria izanda eta programazio ezagutzak ez izatean, hasiera batean honekin lan egitea zaila izatea suertatu daiteke. Hau ekiditeko, gomendio batzuk emango ditugu, hasi berri den erabiltzaile orok argi izan dezan zer egin dezaken azkar garatzeko. Lehenik eta behin, kodea ezagutu beharko dugu. Horretarako, gauza basikoak egiten hasiko gara. Media atera, desbiderazioak, datuak sortu eta hauekin lan errazak egin,…Hau dominatu dugunean, pauso bat aurrera eman beharko dugu. Programazio nozio basikorik gabe ezingo dugu lan konplexurik burutu. Beraz, pauso hau behin eta berriz errepikatu beharko dugu. Era honetan kontzeptu basikoenak automatikoki erabiltzeko ahalmena izango dugu. Honek, denbora asko aurreztuko digu, programarekin lan egitean.

Beste aukera bat, programatzaile garatuenek ere egiten duten gauza bat izango da. Hauek, gehien erabiltzen dituzten komandoak, "Bloc de Notas" programan gordetzen dituzte. Bereziki oso luzeak eta generikoak direnak. Era honetan, denbora aurreztuko dugu, kode guztia ez dugulako idatziko. Gainera, idatzitako kodeak errorerik ez digula emango ziurtatuko dugu. Behin kodea era egokian kopiatu dugun. Era honetan, proiektu luzeekin lan egitean, kode guztia gordeta izango dugu, eta prozesuren bat errepikatu nahi badugu, faktoreak aldatu beharko ditugu bakarrik, lana aurreztuz.

Beste gomendio bat, programatzen ez dakien pertsona batentzat, R pubs web orrian dauden proiektuak ikustea da. Bertan, komandoak nola idazten diren kasu horietarako ikusi dezakegu, eta hortik ideiak atera ditzazkegu gure proiektuan aplikatu ahal izateko. Horrela, nahiz eta programatzen ez jakin, grafiko basikoak egiteko aukera izango dugu. Behin komandoa hartu dugun eta komando honen faktore edo aldagaiak egokitu ditugun.

Beraz, esan dezakegu programazio ezagutzarik gabe programatu daitekeela. Proiektu sinpleetarako nahikoa izango zaigu hau, baino lan konplexuetan beharrezkoa izango da ezagutza minimo at izatea. Horregatik, proiektu guztietan erabiltzen ditugun komandoak era egoki eta ordenatu batean gordetzen baditugu, azkarrago ikasiko dugu programaren funtzionamendua eta gainera, hurrengo lanetarako, komando berdinak erabili behar baditugu, prest izango ditugu, erabili ahal izateko.

R ezagutza baxua duten erabiltzaileentzat, hurrengo gidaliburua sortuko dugu, non komando basiko batzuk agertzen diren, eta histogramen eraketarako eskuragarri ditugun pakete ezberdin batzuk aztertuko ditugu. Gainera, komandoak eskuragarri utziko ditugu, edonork erabili ahal izateko. Ikus dezagun, beraz, R programarentzat prestatu dugun gidaliburua eta bere paketeen azterketa.

4 R SOFTWAREAN HISTOGRAMAK ERATZEKO GIDALIBURUA

Aipatu dugun bezala, berehala ikusiko dugu, lan honen helburua mintzatu duguna. R *softwareak* eskaitzen dizkigun histogramerako aukerak aztertzea, baino lehenago, pausu batzuk ikusi behar ditugu, datuen bilketa, taulen eraketa,…

Grafiko bat irudikatzeko, oinarrizkoa izango da datu batzuk izatea. Datu hauek, R programak dauzkan datuen bankutik, gure dokumentuetatik edo internetetik jeitsi ditzazkegu. Ikus ditzagun aipatutako aukerak:

4.1 DATUAK BILTZEA

Datuak biltzeko aukera ezberdinak ikusiko ditugu. Batzuetan, datuak generatu beharko ditugu, eta beste batzuetan, beste dokumentu batzuetatik bilduko ditugu. R softwareak, datu hauek biltzeko aukerak eskainiko dizkigu. Baita ere, *online* datuak eskainiko dizkigu, gure lan estatistikoak burutzeko. Ikus dezagun ba, R *softwareak* eskaintzen dizkigun aukera hauek:

Datu aleatorioak edo serieak eratu:

- Sekuentzia bat eratu nahi dugunean, adibidez, 1etik 40raino sortuko den sekuentzia numeriko bat, taularaketa bat osatzeko, honako komando hau sartu dezakegu:

```
> seq(1,40)

 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14
15 16 17 18 19 20 21 22 23 24 25 [26] 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40
```

Sekuentzia honek modifikazio batzuk izan ditzazke. Suposa dezagun, datu batzuk errepikatzea nahi dugula, edota datuak beste sekuentzia bat jarraitzea, adibidez, guztiak bakoitiak izatea. Ikus dezagun ba, komando hauek sortzen dituzten datuen adibideak:

```
> seq(1,40,by=2)
```

```
 [1]  1  3  5  7  9 11 13 15 17 19 21 23 25 27
29 31 33 35 37 39
```

Sekuentzia hauek zertarako balio duten irudikatzeko, adibidez, urtean zehar egunero erregistratu diren datuak, denbora serie batekin elkarreko taularaketa bat eratzeko balio ahal izango digu, edota histograma bat eratzerakoan, tarteak jakinda, tarte hauek ze puntutan jarri beharko diren jakiteko. Demagun, tarteak 60 eta 300 arteko zenbakietan egongo direla, eta 7 tarteetan nabarmendu nahi dugula histograma. Komando hau erabilita, tarte bakoitza zein puntutan hasten den jakingo genuke:

```
> seq(60,300, length=7)

[1]      60   100   140   180   220   260   300
```

Datu aleatorioen eraketa:

Datu aleatorioak sortzeko aukera ezberdinak eskainiko dizkigu R *softwareak*. Datuak guztiz aleatorioki ateratzera edota media eta bariantza bat izateko aukera eskeiniko digu programak. Has gaitezen aukera sinpleenekin eta ondoren konplexuagoak direnak ikus ditzazkegu:

Datuak lortzeko aukera bat, *sample* funtzioa erabili dezakegu. Funtzio honek, *n* datu sortuko ditu, datu hauek, bi balioen tartean egoteko edota máximo edo minimo bat izateko aukera izango dugu. Ikus dezagun, 20 datu eskatu ezkero nola ikusiko genuke gure konputagailuan:

```
> sample(1:20,replace=T)

 [1]  6 14 18 17 10 13  9 17 15 16 12 16 12  5
14 17  5 11 11 11
```

Replace T komandoak, datuak errepikatzeko aukera emango digu. *F* jarri ezkero, sortutako datu guztiak ezberdinak izango lirateke.

Beste aukera bat, *Runif* komandoa izan daiteke. Komando hau, *"stats"* paketearen barruan (programa deskargatzean webgune ofizialetik programan

bertan dator[24]) aurkitzen da. *Sample* komandoa bezela, eskatzen ditugun datuak parámetro batzuen tartean egotea eskatu dezakegu. Horrela ikusiko genituzke datuak, programari 60 eta 80 tartean 80 balore aleatorio generatzeko eskatuko bagenio:

```
> runif(80,min=60,max=80)
```

[1] 78.77549 63.78393 74.56984 65.33355 65.38657 77.28118 65.04499 77.46406
[9] 67.25705 65.11398 68.10713 71.24696 79.75817 74.69164 63.97317 71.50515
[17] 69.16722 60.12414 72.30493 62.32625 65.64111 74.15294 75.41623 68.30217
[25] 61.25668 67.06598 76.20924 65.01467 77.11900 70.35258 72.76532 71.85697
[33] 71.94627 73.77369 60.77971 76.75119 73.51287 63.58672 71.15801 68.18150
[41] 73.75862 69.57866 67.53809 77.87687 65.44248 62.85554 77.17364 76.92917
[49] 70.78549 63.12443 63.10435 67.17113 67.86227 74.81796 78.84486 71.70645
[57] 70.95149 73.61104 75.73667 62.99394 76.58453 77.65667 63.38283 71.50032
[65] 70.22468 74.74519 70.70284 61.81749 70.52942 72.93468 63.75115 78.83508
[73] 67.55207 73.29169 78.71425 74.97593 72.28960 77.04685 73.66089 74.01670

Taula 5: R programan runif(80,min=60,max=80) comandotik lortutako datuak. Kontuan izan behar da komando hau errepikatu ezkero, datu ezberdinak lortuko direla, datu aleatorioak sortzen dituelako.

- R datu basetik lortutako datuak

Aipatu dugun bezala, *software* honek daukan abantailetako bat, kode irekiko *softwarea* dela *GNU* lizentziarekin. Horregatik, programarentzat, erabiltzaileek hobekuntzak sortzen dituzte. Honetaz aparte, erabiltzaileek datu baseak argitaratzen dituzte, Rrekin lan egiteko prest, eta edozein erabiltzaileek hartu ditzazke datu hauek, informazioa lantzeko eta ondorioak ateratzeko. Horretarako, komando batzuk sartu beharko ditugu. Lehendabiziko erabili behar dugun komandoa hau izango da:

[24] R 4.0.2 programan automatikoki gordeta egongo da. Izan liteke, programa zaharrago batekin lan egiterakoan, pakete hau deskargatu behar izatea.

```
> data(package=.packages(all.available=TRUE))
```

Komando honekin, pakete guztietan dauzkagun datu pakete guztiak eskuragarri izateko aukera izango dugu, eta hauek eskuratzeko izenak lortuko ditugu. Hau egin ondoren, pakete bat bilatu beharko dugu, lan egiteko. Guk aukeratu duguna *"Car"* paketea izan da. Pakete honen barruan, irakasle batzuen soldaten datu base bat izango dugu, adibidez, soldatarekin zerikusia duten ondorioak ateratzeko. Hurrengo irudian ikusi dezakegu lortuko dugun informazioa.

Irudia 9Estatu batuetako soldataren datuak iturria: egilea Rtik lortuta

Salaries

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
1	Prof	B	19	18	Male	139750
2	Prof	B	20	16	Male	173200
3	AsstProf	B	4	3	Male	79750
4	Prof	B	45	39	Male	115000
5	Prof	B	40	41	Male	141500
6	AssocProf	B	6	6	Male	97000
7	Prof	B	30	23	Male	175000
8	Prof	B	45	45	Male	147765
9	Prof	B	21	20	Male	119250
10	Prof	B	18	18	Female	129000
11	AssocProf	B	12	8	Male	119800
12	AsstProf	B	7	2	Male	79800
13	AsstProf	B	1	1	Male	77700
14	AsstProf	B	2	0	Male	78000
15	Prof	B	20	18	Male	104800
16	Prof	B	12	3	Male	117150
17	Prof	B	19	20	Male	101000
18	Prof	A	38	34	Male	103450
19	Prof	A	37	23	Male	124750
20	Prof	A	39	36	Female	137000
21	Prof	A	31	26	Male	89565
22	Prof	A	36	31	Male	102580
23	Prof	A	34	30	Male	93904
24	Prof	A	24	19	Male	113068

Pakete honetaz aparte, pakete ezberdinak aukeratu ditzazkegu datu hauekin ondorioak ateratzeko. Adibidez, *Titanic* itsasontziaren bidaiari guztien datuak deskargatu daitezke. Bertan, zeintzuk salbatu ziren eta zeintzuk ez agertzen da, ze bidaiari clase ziren, adina, sexua,…Azken finean, programa lantzeko aproposa izan daiteke hasiera batean, programak berak eskaitzen dituen liburutegitik informazioa ateratzea, eta bertako informazioarekin ariketak egitea programaren funtzionamendua ikasteko.

- Rtik kanpo aurkitu daitezkeen datu baseak

Eboluzio teknologikoak, datuak era errazago batean biltzeko aukerak eman dizkigu, eta honekin batera, internetaren laguntzarekin,

datuak partekatzeko aukera infinituak eskaini dizkigu. Gaur egun, webgune askok datu base ezberdinak eskaintzen dituzte. Horrela, gure eskura aurkitu dezakegun informazioa, orain dela 30 urte baino askoz erabilgarriagoa suertatuko da. Datu base ezberdin hauek aurkitzeko, garrantzitsua izango da bilaketa ingelesez egitea (*Datasets, Google* bilatzailean ipini). Normalean, datu base hauek *Excel* formatoan eskainiko dituzte, beraz, *Excel* formatotik Rra pasatzeko gai izan beharko gara (Hurrengo atalean ikasiko dugu).

Aipatu dugun bezala, datu base ezberdinak dituzten webgune ezberdin asko daude. Erabiliena, *Kaggle* webgunea izango da. Webgune hau, erabiltzaileek garatuko duten webgunea izango da. Hauek igoko dituzte datu base ezberdinak, besteek doan erabili ahal izateko. Garrantzitsua izango da, informazio hau, edonork igo dezakeela, eta datu base bat informazio ez erabilgarri batekin lotu daitekeela. Garrantzia eman beharko diogu, beraz, informazioa webgune ofizialetatik lortzea, baino egongo dira kasuak, non informazio hau ez den egongo era ofizialean, beraz, portal hauen laguntza behar izango dugu.

- Datuak ordenagailutik inportatzeko aukera

Lehen aipatu dugun bezala, programa honek datuak sartzeko aukera ugari eskaintzen dizkigu. Aukera hauetako bat, gure dokumentuetatik eratorritako informazioa sartzea izango da. Aukera hau, izugarrizko abantaila bat izango da erabiltzailearentzat, datu askoko informazioa era azkar batean prozesatu ahalko dugulako. Adibidez, Euskal Herriko informazioa aztertu nahi dugunean, *EUSTAT* orrialdean sartu ezkero, informazio ugari lortu dezakegu, baino informazio hau guztia prozesatzeko erramintarik gabe, ez da oso erabilgarria izango. Ikus dezagun nola pasa dezakegun R programara informazio bat, *EUSTAT* webgunetik atera duguna:

Lehenik eta behin, *Eustat* web orrialdean sartu gara eta aktualitateko tema bat aukeratzea erabaki dut. Oso aipatua da azkeneko urtetan, etxebizitzen prezioak izan duen garapena, eta *Euskal Herriko estatistika institutuak*, prezio hau ze faktoreengatik igo den aztertzen du. Faktore hauen artean, materialen

prezioa izango da, beste faktore bat, eskulanaren kostua izango da, obra zibilaren kostua,…Aldagai hauek guztiak, R programan automatikoki barneratzeko aukera izango dugu. Noski, *EUSTAT*etik lortutako informazioa *(etxebizitzaren kostuaren garapena 2016 urtetik Aurrera, hileroko garapena aztertuz*[25]*) Excel* dokumentuan gordetzea gomendatuko dugu. Lehenago aipatu dugun bezala, informazio guztia era egokian izango baitugulako taularatuta eta *softwarean* barneratzeko erraztasun haundiagoa izango dugu. Hau esanda, programan, dokumentu batetik informazioa barneratzeko bi aukera izango ditugu, segituan ikusiko ditugunak, baino lehenago ikus dezagun *EUSTAT*etik lortu dugun dokumentuak zein informazio eskaintzen digun:

Irudia 10Etxebizitzaren kostuaren garapena 2016 Iturria: EUSTAT

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1			2016-01	2016-02	2016-03	2016-04	2016-05	2016-06	2016-07	2016-08	2016-09	2016-10	2016-11	2016-12	2017-01	2017-02	2017-03
2	Costes total	Índice	102,5	101,8	101,7	101,3	101,1	101,2	101	101,2	101,5	102,1	101,8	101,8	102,6	102,6	102,4
3		Incremento sobre mes anterior	0,7	-0,7	-0,1	-0,5	-0,2	0,1	-0,2	0,2	0,3	0,6	-0,3	0	0,8	0	-0,2
4		Incremento sobre mismo mes año anterior	0,6	-0,2	-0,5	-0,7	-0,9	-0,5	-1	-0,9	-0,3	0,6	0,3	0,1	0,1	0,8	0,6
5		Incremento sobre diciembre anterior	0,7	0,1	0	-0,5	-0,6	-0,5	-0,7	-0,6	-0,3	0,4	0,1	0,1	0,8	0,8	0,5
6	-Costes edificación	Índice	103	102,3	102,2	101,7	101,4	101,5	101,3	101,5	101,8	102,3	102	101,9	102,7	102,7	102,5
7		Incremento sobre mes anterior	0,9	-0,7	-0,1	-0,5	-0,3	0,1	-0,2	0,1	0,3	0,6	-0,3	-0,1	0,7	0	-0,2
8		Incremento sobre mismo mes año anterior	0,9	0,2	-0,1	-0,3	-0,6	-0,2	-0,7	-0,8	-0,2	0,6	0,2	-0,1	-0,3	0,4	0,3
9		Incremento sobre diciembre anterior	0,9	0,2	0,1	-0,3	-0,6	-0,5	-0,7	-0,5	-0,3	0,3	0	-0,1	0,7	0,7	0,5
10	-Costes obra civil	Índice	99,6	99	99,2	98,8	99	99,4	99,2	99,4	99,8	100,8	100,5	101,1	102,1	102,2	101,6
11		Incremento sobre mes anterior	-0,4	-0,6	0,2	-0,4	0,3	0,4	-0,3	0,2	0,4	1	-0,4	0,6	1	0	-0,5
12		Incremento sobre mismo mes año anterior	-1,3	-2,5	-2,8	-3,2	-3,1	-2,3	-2,5	-1,8	-0,8	0,7	0,4	1,1	2,5	3,2	2,5
13		Incremento sobre diciembre anterior	-0,4	-1	-0,8	-1,3	-1	-0,6	-0,8	-0,6	-0,2	0,8	0,5	1,1	1	1,1	0,6
14	Precios de materias primas total	Índice	101,4	100,5	100,4	99,8	99,6	99,8	99,5	99,7	100,1	100,9	100,5	100,5	101,1	101,1	100,8
15		Incremento sobre mes anterior	0,7	-0,8	-0,1	-0,6	-0,2	0,2	-0,3	0,2	0,4	0,8	-0,4	0	0,6	0	-0,3
16		Incremento sobre mismo mes año anterior	0,6	-0,4	-0,8	-1,2	-1,4	-0,9	-1,5	-1,4	-0,6	0,6	0,1	-0,1	-0,3	0,6	0,4
17		Incremento sobre diciembre anterior	0,7	-0,1	-0,2	-0,8	-1	-0,9	-1,1	-0,9	-0,5	0,3	-0,1	-0,1	0,6	0,6	0,3
18	-Precios de materias primas edificación	Índice	102	101,1	101	100,4	100	100,2	99,9	100,1	100,5	101,2	100,8	100,7	101,2	101,2	101
19		Incremento sobre mes anterior	1	-0,9	-0,2	-0,6	-0,3	0,1	-0,3	0,2	0,4	0,7	-0,4	-0,1	0,5	0	-0,3
20		Incremento sobre mismo mes año anterior	1	0,1	-0,3	-0,6	-0,9	-0,5	-1,2	-1,2	-0,5	0,6	0,1	-0,3	-0,8	0,1	0
21		Incremento sobre diciembre anterior	1	0,1	0	-0,6	-0,9	-0,8	-1,1	-0,9	-0,5	0,2	-0,2	-0,3	0,5	0,5	0,3
22	-Precios de materias primas obra civil	Índice	97,6	96,8	97,1	96,5	96,8	97,4	97	97,3	97,9	99,2	98,7	99,5	100,4	100,5	99,8
23		Incremento sobre mes anterior	-0,7	-0,8	0,2	-0,6	0,3	0,6	-0,3	0,3	0,6	1,4	-0,5	0,8	0,9	0	-0,7
24		Incremento sobre mismo mes año anterior	-1,9	-3,4	-3,8	-4,4	-4,3	-3,2	-3,5	-2,6	-1,3	0,7	0,3	1,2	2,9	3,8	2,8
25		Incremento sobre diciembre anterior	-0,7	-1,5	-1,3	-1,9	-1,5	-1	-1,3	-1	-0,5	0,9	0,4	1,2	0,9	1	0,3

Ikusi dugun taula hau, *Excel* dokumentu batetik R dokumentu batera pasatzeko, lehenengo *Excel* dokumentua prestatu behar dugu. Dokumentuak, zenbat eta hutsune gutxiago izan, hobeto irakurriko ditu. Hau da, izenburua kendu beharko dugu, eta aldagai funtzioa izango dituzten zutabeak

[25] EUSTAT (2020). Eustat.eus. Iturria: http://es.eustat.eus/bankupx/pxweb/es/spanish/-/PX_2292_icce1m_01.px#axzz5m5jx4mq7 etxebizitzaren kostuaren garapena 2016 urtetik Aurrera, hileroko garapena aztertuz

lehendabiziko zutabe eta lerroan ipintzea gomendagarria izango da. Beraz, datu guztiak batera gure R kontsolan barneratzeko komandoa honako hau izango lirateke:

```
>readXL("C:/Users/usuario/Desktop/icce1m_01(1)
.xlsx",rownames=TRUE,header=TRUE,na="",sheet="icce1
m_01", stringsAsFactors=TRUE)
```

- Informazioa gure kabuz prestatzea

Demagun, datu basea gure kabuz sortu nahi dugula. Honek, denbora asko kenduko digu, eta gainera, jarraipuntu batzuk jarraitu beharko ditugu, datu base hau R programarekin bat etortzeko. *STHDA* web horriak[26] laguntza eskaintzen digu, bide hau jarraitu nahi badugu.

Beraz, adituek diotenez, zutabeak, aldagaiak izango diren baloreak errepresentatzea lagungarria izango da, eta lerroetan, izenak, obserbazioak edota analizatuko ditugun datuak barneratuko ditugu. Garrantzitsua izango da, datu hauen artean, izenik ez duplikatzea, Rk datuak hobeto irakurri ahal izateko.

Modu honetan informazio guztia barneratuko dugu programan bertan eta informazio guztia garatzeko aukera izango dugu, baina zer gertatuko lirateke, dokumentu osoa R *softwarean* barneratu ez nahi badugu, eta bakarrik dokumentu horren zati, zutabe edo lerro bat lortu nahi badugu? Suposa dezagun, ikertzaile batek etxebizitzen kostu totalaren eboluzioa aztertzen duela, eta kostu honen hazkuntza zenbatestu nahi duela histograma bat eratuz. Horrela, azkeneko urtetan ikusi dezake hazkundea nabariagoa izan dela murrizketak baino, edota tendentzia bat dagoela aztertu nahi du. Horretarako, dokumentuaren A5 lerroarekin nahikoa izango du. Komenigarria izango da dokumentua prestatzea, nahi dugun informazioa barneratzeko, bestela, informazio gehiegi izan ezkero, programazio maila baxua izan ezkero, grafiko zehatzak edo zuzenak eratzeko zailtasun

[26] http://www.sthda.com/english/wiki/best-practices-in-preparing-data-files-for-importing-into-r linkean ikusi daiteke, estatistikan espezializtutako web orrialde honek, Rrekin lan asko egiten duena, zeintzuk dira jarraipuntuak.

haundiagoak izango ditugu. Beraz, dokumentua importatu baino lehen, honen kopia bat sortuko dugu, nahi dugun informazioa barneratuz bakarrik.

Ez dugu lehenago aipatu, baina garrantzitsua izango da, ikertzaileak jasoko duen informazioa webgune ofizialetatik jasotzea. Datuak biltzeko garrantzitsua izango da noski, taularaketa egokia izatea, baino hau bezainbeste garrantzitsuagoa, edo gehiago, gure informazio iturria iturri ofiziala izatea izango da. Beraz, nahiz eta datu bilketan, R konputazioan zerikusirik ez izan, datuen tratamenduan garrantzi haundia izango duen faktorea izango baita.

Lanaren zati hau esplikatzen bukatzeko, R programarekin lan egiteko, datuak banaka sartzeko aukera izango dugu. Era honetan, guk nahi ditugun datu zehatzak barneratzeko aukera izango dugu. Ez da komenigarria izango, datu kopuru asko izan ezkero, eredu hau jarraitzea gure lana garatzeko, izugarrizko denbora galera suposatuko duelako, beste edozein erarekin konparatuta. Era labur batean ikusiko dugu, datuak barneratzeko metodoa zein den, 5 pertsonen altuera (zentimetroetan) eta pisua (kilogramotan) ipiniz:

```
> pisua=c(80,73,60,105,75)

> altuera=c(185,168,164,198,181)
```

Hau esanda, datu guztiak gure R komputagailuan barneratu ezkero, histograma eratzeko pausoak jarraitu beharko ditugu. Horretarako, R programatik ateratako datu base bat aukeratzea erabaki dut (horrela, datuak prest daude hauen tratamendurako), lan honen helburua, R *softwareak* eskaintzen dizkigun aukerak histograma bat eratzean aztertzea baitelako. Hau aztertu ondoren, noski, kasu praktiko batean testatuko dugu, baino garrantzitsua izango da teoria kontzeptuak ulertzea, kasu praktikoa ondo ulertzeko.

R datu baseen erabilgarritasuna dela eta, hemen ikusiko ditugu datu base batzuk. Hauetako batzuk ondorengo kasu praktikoetan erabiliko ditugu adibide bezela. R plataformak bere web orrialdean eskaintzen ditu eta edozein erabiltzailearentzako erabilgarri izango dira.

4.2 Read R paketea

Lehen aipatu dugun bezala, R *softwarea* aurkeztean, erabiltzaileek hauen garapenak partekatzeko ohitura zutela adierazten genuen. Ondoren ikusiko ditugu grafikoak sortzeko pakete ezberdinak, baino datuak irakurtzeko ahalmena eskaintzen dizkiguten paketeak ere eskuragarri izango ditugu. *Read R* paketea orretarako sortu zen. Erabiltzaile hasiberrientzat datuak barneratzeko erraztasunak eskaintzeko. *Tidyverse* erabiltzaileak eraikitako paketea da (GGPLOT2 paketearen sortzailea baita ere, ondoren ikusiko duguna) aztertuko dugun paketea. Bertan, funtzio ezberdinak ikusiko ditugu datuen tratamendurako. *Readr* funtzioak, artxiboak data frametan bihurtzeko helburua izango dute:

- Read.csv(): Koma batekin aldenduta dauden artxiboen irakurketa lortuko du. Hau da, 1,2,3 eta 4 datuak, koma batekin aldenduta egon ezkero (1,2,3,4) irakurtzeko gai aproposa izango da. *Read_csv2()* funtzioak berriz, berdina egingo du, kasu honetan, puntuak erabiliz elementuak tartekatzeko. Bi aukera dauzkagu, herrialde batzuetan zenbaki dezimalak puntu baten ondoren jartzen direlako, eta beste batzuetan koma baten ondoren. *Read.tsv()* tabulazioengatik delimitatutako elementuak irakurriko ditu eta *read_tsv()* edozein delimitadorekin irakurriko ditu datuak. Honek esan nahiko du, puntu eta komak izanez gero, azkeneko kasua ez dela komenigarria izango gure datu basea osatzeko.
- *Read_fwf()* zabalera fijo baten artxiboak irakurtzen ditu. Zabaleragatik irakurtzea ez nahi badugu, komando ezberdinak erabiltzeko aukera izango dugu *fwf_widths()* bere ubikazioagatik aukeratuko du adibidez. Beste kasu bat *fwf.positions().read.table()* izan daiteke. Komando honek zabalera ezberdineko kolumna ezberdineko datuak barneratzeko aukera eskainiko digu.

4.3 Datuen prestakuntza

Datuak iturri ezberdinetatik ateratzen ditugunean, askotan gertatuko zaigu gure erabilpenerako ez dela formatu egokian egongo. Paketeekin jarraitu baino lehenago, jakinarazi dezagun, datu base honekin lan egingo dugula hurrengo ataletan. Beraz aztertu dezagun datu base honek ekartzen duen informazioa, egiten ditugun komandoak ulertzeko. 1936 urtean egin zen ikerketa batetik eratorritako

datu basea izango da. Bertan, 3 landare espezie desberdinekin ikerketa batzuk egin zituzten. Bere petaloen lodiera eta luzera neurtu zituzten, eta gainera, sepaloekin berdina egin zuten. Datu base txiki bat denez, programak azkarrago lan egingo du. Ez da izango komenigarria izugarrizko datu baseekin lan egitea, horretarako prest ez dagoen ordenagailurik ez badugu. Gertatu daiteke komando bat eskatzean, komando hori burutzeko 2 ordu edo gehiago irautea.

Adibidez, elementu baten izena era desegokian egon daiteke. Imagina dezagun *iris* datu basearekin ari garela lanean, eta programari komando ezberdinak sartzeko intentzioa dugula. Datu base honen elementuetako bat, *Petal.width* izango da, konplexua idazten. Elementuen edo *header* funtzioa duten elementuak, aldatzeko aukera izango dugu, horrela, programatzen hasterakoan, izenburuak gure gustura izateko. Era honetan, erabiltzailearentzat erraztasun haundiago bat izatera bultzatuko du. Demagun, datu base honekin lanean ari garela, eta *petal.width* haundienak identifikatu nahi ditugula. Horretarako, *iris[order(-iris$Petal.Length),]* komandoa erabili beharko dugu.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
119	7.7	26	6.9	2.3	virginica
118	7.7	3.8	6.7	2.2	virginica
123	7.7	2.8	6.7	2.0	virginica
106	7.6	3.0	6.6	2.1	virginica
132	7.9	3.8	6.4	2.0	virginica
108	7.3	2.9	6.3	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
131	7.4	2.8	6.1	1.9	virginica
136	7.7	3.0	6.1	2.3	virginica
101	6.3	3.3	6.0	2.5	virginica

Hauek izango dira lehenengo 10 datuak. Antzemango dugu *virginica* espezieak petalo lodiera haundiagoa izango duela, beste espezieak baino. Ondorio hau, datu guztiak ondo ordenatuta baldin badaude, era errazago batean antzemateko aukera izango dugu. Horregatik erreparatuko dugu, prestakuntza lanaren garrantzian. Begibistan pasa daitezkeen datuak, azkarrago identifikatuko ditugu era honetan.

Gainera, datu base batean elementu kualitatiboak daudenean, eta gure datuen interpretaziorako elementu kualitatibo hauen artean filtro bat erabili behar badugu, kortxeteak ([ ]) erabiliko ditugu. Era honetan, programari esango diogu, gure datu guztietatik zein izango da interesatzen zaiguna. Demagun *iris* datu basetik, *setosa* espeziea bakarrik ikertu nahi dugula. Horretarako, filtro komando bat erabili beharko dugu, hau izango dena: ***>iris[iris$Species == "setosa",].*** Era honetan, hiru espezie mota ezberdinen datu baseak filtratzeko gai izango gara, eta espeziearekiko ikerketak errazagoak izango dira. Beraz, hiru komando hauek erabiliko ditugu, datuak ordenatzeko:

```
>iris[iris$Species == "versicolor",].

>iris[iris$Species == "setosa",].

>iris[iris$Species == "virginica",].
```

Era honetan, jasotako informazioa era ordenatuago batean izango dugu, lehen esan bezala. Beraz, datuekin lan egiteko erraztasun haundiagoak izango ditugu modu erraz batean. Aipatutako bi aukerak, konbinatzeko aukera izango ditugu. Adibidez, *Setosa* espeziea petalo lodieraren arabera ordenatzeko aukera izango dugu

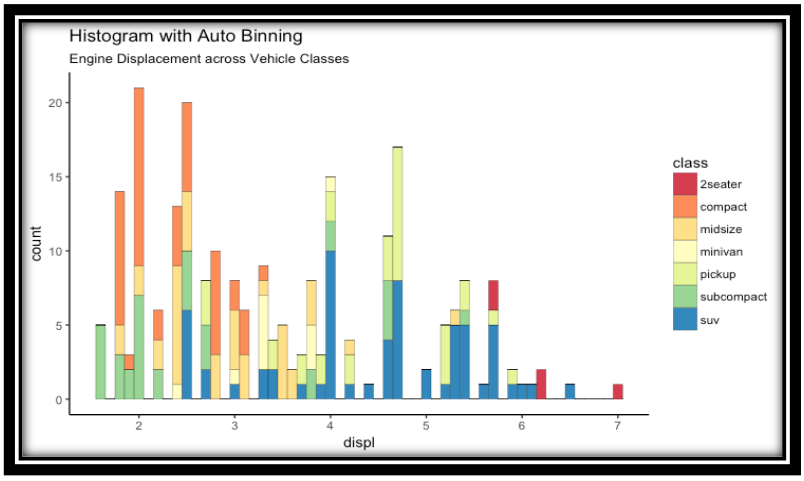
4.4 TAULARAKETA ETA GRAFIKOAREN SORRERA

Dakigunez, grafiko bat sortu baino lehen, komenigarria da datuen taularaketa zuzena egitea. Programarekin taularaketak burutzeko aukera desberdinak izango ditugu, eta automatikoki kalkulu ezberdinak egitea eskatu diezaiokegu. Baita ere, konputazioan sortzen diren problemak ekiditeko, gure tauletan dauden balio kualitatiboei, balio kuantitatibo batzuk eransteko aukera izango dugu.

Demagun gure datuak hartzen ditugun datu basetik, balio kualitatibo bat dagoela, gizona balioa izango duena sujetua gizona denean, eta emakumea balioa izango duena, sujetua emakumea denean. Honek, kalkulu batzuk lortzeko, problema bat suposatuko du, programak zenbakiekin lan egingo duelako gehienbat. Horregatik, bietako balio bati 0 balioa emango zaio, eta besteari 1 balioa. Hau da, gizonak diren sujetuak, 1 zenbakia izango dira, eta emakumeak, berriz 0 balioa edo alderantziz. Horrela, programak, gizonen artean dagoen media, emakumeena, bi grafiko ezberdinak konparatzeko edota guztiak batera elkartzeko aukera izango du.

Behin gure datuak ondo sartu ditugun, garrantzitsua izango da zein histograma mota garatu nahi dugun. Biztanleria piramide bat sortzeko adibidez, *"Psych"* paketea deskargatu beharko dugu. Programan bertan, grafikoak garatzeko pakete ezberdinak aurkitu ditzazkegu. Hauetako bat, *"GGPLOT2"* paketea izango da, grafikoetan espezializatua izango dena. R paketeetan, erabiltzen den pakete nabariena izango da. Horrelako grafikoak[27] sortzeko aukera emango digu adibidez:

Irudia 11Kotxe kopuruen eta hauen arteko desberdintasunak. Iturria: egilea



Histogramaren eraketarako, lanaren lehendabiziko atalean ikusi dugun bezala, garrantzitsua izango da tarteak definitzea. Programak, *Sturgesen metodoa* erabiliko du automatikoki, beraz, komando espezialak erabili beharko ditugu beste metodoren bat erabiltzeko, edota guk nahi dugun bezala

[27] http://r-statistics.co/Top50-GGPLOT2-Visualizations-MasterList-R-Code.html#Histogram web orrialdean ikusi daitezke GGPLOT2 paketearekin sortu daitezkeen grafiko mota nabariak. Bertatik, histograma hau lortu dugu, non clase ezberdinak aldentzen dituen histograma sortzen du.

tartekatzeko. Eskuz sartu nahi baditugu tarteak komandoen bidez, honako komandoa erabili beharko dugu:

```
>hist(pertsonak,breaks=c(0,10,20,30,40,50,60
,70,80,90,100,250)
```

Ikusi daitekeenez, komandoko azken zenbakia tarte normalak baino haundiagoa jarri dugu, batzuetan, grafika batek horrelako tarteak sortzeko obligazioa sortzen dutelako. Kasu honetan, *softwareak* ez digu automatikoki tarte irregular bat sortuko, beraz, komenigarria izango da komando bidez sartzea tarteen lodiera, tarte irregularren histograma bate egin nahi bada. Metodoren bat erabiltzeko, erabili beharko diren komandoak hauek izango dira:

Taula 6Metodoen laburdurak Rn barneratzeko. Iturria: egilea

METODOA	KOMANDOA
STURGES	*Breaks=”sturges”*
FREEDMAN DIACONIS	*Breaks”fd”*
SCOTT	*Breaks=”Scott”*

Beraz, behin histograma eta bere tarteak definitu ditugun, garrantzitsua izango da, bertan ematen den informazioa definitzea, irakurle edo ikertzaileari informazioa errazteko. Estatistika deskribatzailearen funtzioa hauxe bera da, zenbat eta informazio gehiago barneratu grafikoan hobe, ulergarria baldin bada.

Adibidez, izenburua aldatzeko, *main=(nahi dugun izenburua)* komandoa erabili beharko dugu. Gainera, azpititulua gehitzeko aukera izango dugu. Honetaz aparte, *x.lab* eta *y.lab* gure grafikoen ejeetako etiketen izena pertsonalizatzeko aukera emango digute. Ejeen lodiera aldatzeko aukera izango dugu, eta honen zabalera aldatzeko aukera ere. Hurrengo laburpenean ikusi ditzazkegu R pakete basikoak ekartzen dituen aukerak histogramak hobetzeko:

- *main*=Izenburua zehazteko erabiliko dugun komandoa *(main="nahi duguna")*.
- *sub*=Azpititulua zehazteko erabiliko dugun komandoa *(sub="nahi duguna").*
- *type*=Grafiko marra mota zehazteko erabiliko dugun komandoa, puntuzko barrak, adibidez, "dashed" komandoarekin zehaztuko ditugu.
- *lty="1"* Marra mota
- *pch="."* Marrazki mota
- *xlab=X* ardatzaren izenburua aldatzeko zehaztuko dugun komandoa *(xlab="nahi duguna").*
- *Ylab= Y* ardatzaren izenburua aldatzeko zehaztuko dugun komandoa *(ylab="nahi duguna")*.
- *xlim=c(xmin,xmax)* X ardatzaren eskala definitzeko aukera izango dugu, grafikoa ondo ajustatzeko gure beharretara *(xlim=c(xmin=0,xmax=10))* komandoa erabili beharko genuke, X ardatza 0tik 10arte iristea nahiko bagenu.
- *ylim=c(ymin,ymax)* Y ardatzaren eskala definitzeko aukera izango dugu, grafikoa ondo ajustatzeko gure beharretara *(ylim=c(ymin=0,ymax=10))* komandoa erabili beharko genuke, Y ardatza 0tik 10arte iristea nahiko bagenu.
- *col="456"* koloreen eskala aldatzeko aukera emango digun komandoa. Garrantzitsua izango da lan honetan ateratako ondorioak, aurkezteko garbi egotea, beraz, ondoren, azpiatal bat sortuko da, koloreekin egon daitezkeen funtzio guztiak aztertzeko.

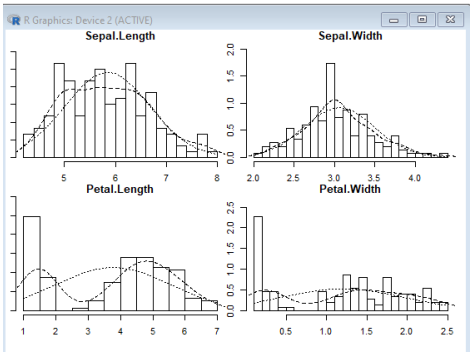
Pakete basikoaz aparte, R programak pakete ezberdinak eskainiko dizkigu, gure histogramaren garapenean laguntzeko. Behin histograma bat eratzeko komando basikoak ulertu ditugula, paketeak ezagutzea gomendagarria izango da. Paketeak programa edo softwarearen gehigarri bat dela ulertu behar dugu, eta pakete basikoa dominatu gabe honek eskaintzen dizkigun extrak lantzea zailagoa iruditu zaigu. Etxea teilatutik eraikitzen

hastea bezala izango da. Hurrengo azpiataletan, programa hauek aipatu eta aztertuko ditugu.

4.4.1 Psych paketea

Psych paketea oso erabilgarria izango da, gure azterketaren helburua histograma bat baino gehiago aztertzea denean eta hauen artean konparaketak egin daitezkeenean. Programa honekin, histogramak bateratzeko aukera izango dugu. Horrela, adibidez, biztanleria piramide bat sortzeko aukera izango dugu, edota datu kualitatibo ezberdinen maiztasunak konparatzeko ahalmena izango dugu, histograma desberdinak eratuz eta batera elkartuz. Datozen aukerekin, *Psych* paketeak datu base ezberdinak ekarriko ditu. Behin paketea deskargatu dugula eta erabilgarri jarri dugunean *(library(psych))*, *Iris* izeneko datu basearekin lan egingo dugu(Pakete guztien esplikazioetan, pakete berdina erabiltzen saiatuko gara). Horretarako, *multi.hist* funtzioa erabiliko dugu. Funtzio honek, datu baseko aldagai bakoitzeko histograma bat eratuko digu automatikoki. Kasu honetan, 4 aldagai daudenez, 4 histograma eratuko ditu batera. Honek, lana aurreztuko digu, eta grafika guztiak batera ikusteko aukera izango dugu, emaitzak bilatzeko aukera errazagoa emango duguna. Ondoren ikusi dezakegu emaitza nolakoa izango den:

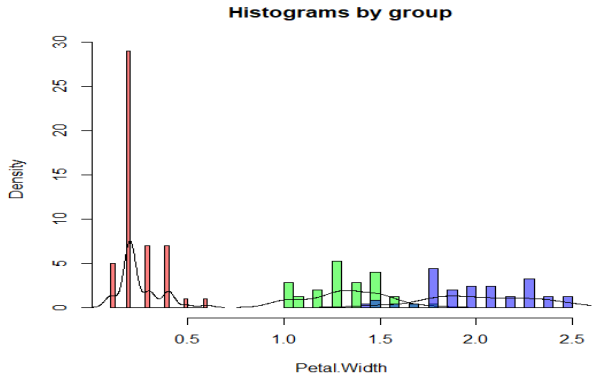
Irudia 12Iris datu besearen aldagaien histograma Rrekin landuta. Iturrria: egilea



Ez da izango dugun aukera bakarra. Pakete honen ekarpena, esan dugun bezala, ez da bakarrik histograma ezberdinak batera eratzea, baizik eta histograma batean, aukera ezberdinak bateratzea grafika berdin batean egiteko. Kasu honetan, erabili beharko dugun komandoa ezberdina izango da. *HistBy* komandoak, aukera hau emango digu, adibidez, genero ezberdinak aztertzea nahi badugu eta *SATQ* aldagaia histograma batean eratu nahi

badugu, sartu beharko dugun komandoa hau izango da; *histBy(iris,"petal.Width","species")*Emaitza honako hau izango da:

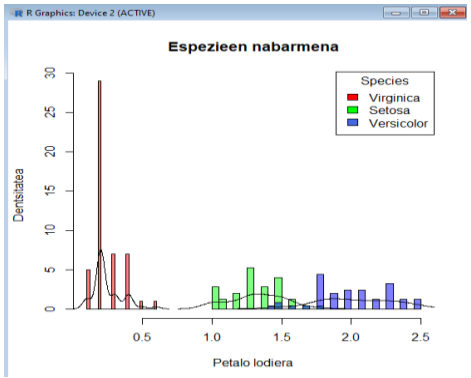
Irudia 13Espezieen arteko desberdintasuna aztertzeko histograma Rrekin landuta. Iturria: egilea



Grafiko honetan, argi ikusiko da, espezie bakoitzak petalo lodiera ezberdin bat izango duela. Garrantzitsua izango da, histograma, edozein erabiltzailek ulertzeko aproposa izatea, beraz, legenda bat erabiltzea eta izenburuak editatzea komenigarria izango da. Horretarako, gogora dezagun *xlab, ylab, Main* eta *leyenda*, orain esplikatuko dugula zein komando erabili behar diren, leyenda egokia ipintzeko:

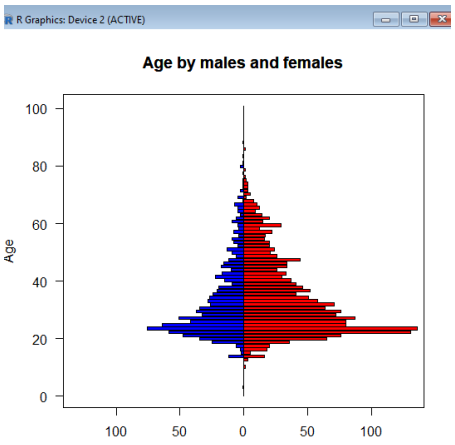
legend(x = "topright", legend = c("Virginica", "Setosa","Versicolor"), fill = c("red", "green","royalblue"), title = "Species") komandoa erabiliko dugu. *X=topright* arauak, leyendaren kokapena zehaztuko du. Normalean, grafikoetan kokapen hau izango da. Ondoren, *legend = c("Virginica", "Setosa","Versicolor")*, arauak, legendak izango dituen aukerak izango dira. Hurrengo komandoak koloreak definitzeko baliagarri izango zaizkigu. Ikus dezagun, beraz, nola geratuko da gure grafika.

Irudia 14 Irudia, irakurlearentzat egokituta. Iturria: Egilea



Askotan komentatu dugun bezala, beharrezkoa izango da grafikoaren egitura, edonork ulertu dezan prest egotea. Ikusi daiteke, bi grafiken arteko desberdintasuna nabaria dela, nahiz eta datu berdinak izan. Lehenengo grafikatik ondorio gutxiago aterako ditu grafikoaren irakurleak, denbora berdinean aztertuko balitu bi grafikak. Horregatik, grafikaren helburua betetzea lortu behar dugu. Ahal den azkarren barneratu beharko du irakurleak ahal den informazio gehiena.

Pakete honekin bukatzeko, aipatu dugu biztanle piramidea burutzeko aukera izango dugula. Horretarako, beste datu base bat erabili beharko dugu, *bfiren*[28] kasua izango dena. Datu base hau erabiliko dugu era labur batean, gizon emakumeen ñabardura egiteko histogramaren eraketan. Horretarako, komando espezial bat erabiliko dugu, *bi.bars* izango dena. Beraz, *Bfi* paketea deskargatu ondoren, komando hauek erabiliko ditugu, biztanleria piramide bat eratzeko, gizon eta emakumeak alderatuz.



Irudia 15Gizon emakumeen biztanleria piramidea Rrekin landuta. Iturria: egilea

Beraz, ikusi dugunez, *Psych* paketeak, datu base ezberdinak eskainiko dizkigu grafikekin lan egiteko, eta histogrametan egin daitezkeen aukerak nabariak dira. Alde batetik, *Multi.hist* komandoak, datu base batetik aldagai guztien histogramak ateratzeko aukera emango digu. Horrela, histogramen eta hauen dentsitate funtzioen konparaketa zuzena egin daiteke denbora labur batean. Bestetik, aldagai bat nabarmendu nahi dugunean, gure kasuan adibidez landare espezieen artean dauden desberdintasunak aztertzeko, *Hist.by* funtzioa erabili dezakegu. Funtzio honek, espezie bakoitzak izango duen petalo

[28] Bfi datu baseak, 28 aldagai izango dituen datu basea izango da. Bertan, sexua, eta adinaz aparte, pertsonalitatea aztertzen dituzten galderen erantzunak barneratuko dira. 2800 pertsona baino gehiago ikertu ziren lan honetarako. Emaitzak eta informazio gehiago, hurrengo webgunean ikusi ahal izango dira: https://ipip.ori.org/

lodieraren histograma marraztuko digu. Horrela konturatuko gara, espezie bakoitzak lodiera desberdin bat izango duela. Bukatzeko, nahiz eta gure datu basearekin komando hau erabilgarria ez izan, ikusi daiteke pakete honek biztanleria piramideak eratzeko aproposa dela. Beraz, grafikak burutzeko, pakete hau eskuragarri izan beharko dugu. Pakete hau, ez da bakarra izango, gure R *softwarean* izan beharko duguna. Pakete erabilgarri asko egongo dira eskuragarri, *Lattice* paketearen kasua bezala adibidez. Orain, *Lattice* paketeak izango dituen aukerak aztertuko ditugu.

4.4.2 Lattice paketea

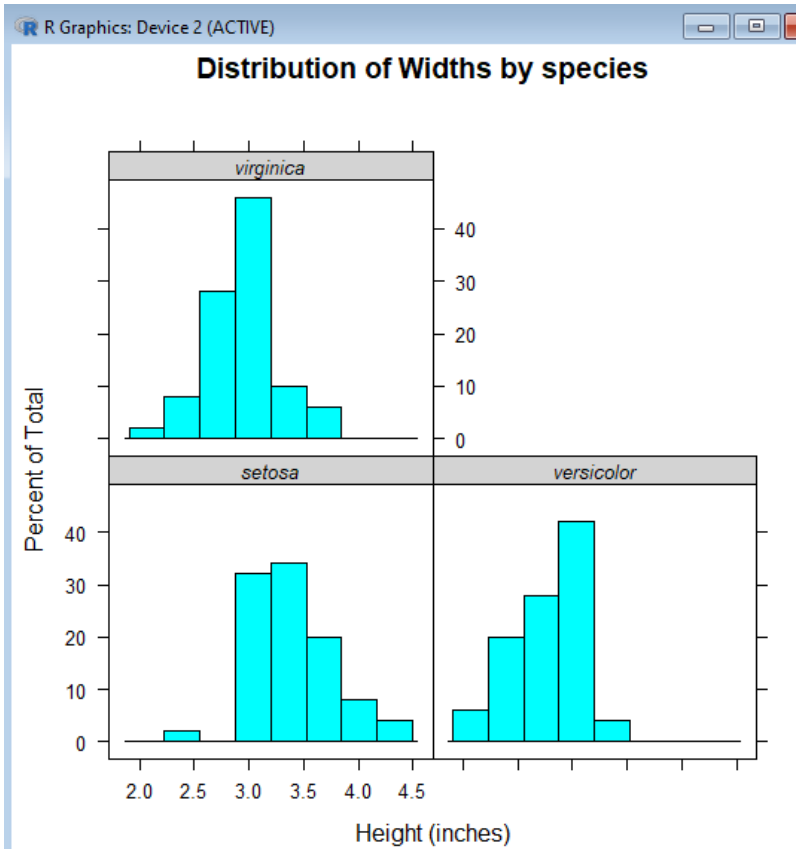
Lattice paketea, seriez instalatutako pakete bat izango da, R *softwarearen* barnean. Honen erabilerarako bakarrik liburutegitik hartu beharko dugu *library* komandoa erabiliz. Erabilera ezberdinak izango ditugu pakete honekin histograma baten eraketan, ondoren ikusiko dugun bezala, aldagai batzuk nabaritu egingo ditzazkegu eta histograma ezberdinak eratzeko gaitasuna izango dugu. Horrela, datuak klasifikatzeko eta klasifikazio honen emaitzekin histogramak eratzeko aukera izango dugu.

Jarrai dezagun gure lanarekin eta azter dezagun *Lattice* paketeak ekar ditzazkeen hobekuntzak. Lehen komentatu dugu, *Psych* paketeak espezieak desberdintzeko aukera ematen zigula. Ez da pakete bakarra izango, funtzio hau bete dezakeena. Esan dugunez, R sistema ireki bat denez, erabiltzaile askok pentsatu dute, programak grafika ezberdinak batera erakustea komenigarria izan daitekeela, horregatik, pakete askotan aukera hauek ikusiko ditugu. Jarrai dezagun beraz, *Iris* datu basea lantzen. Beraz, lehen bezala, espezieen arteko desberdintasunak ikusteko, desberdintasun hau egingo dugu. Kasu honetan, *setaloen* arteko desberdintasuna aztertuko dugu.

Histogramak bateratzeaz aparte, beste grafiko motak bateratzeko aukera ezberdinak eskainiko dizkigu pakete honek. Adibidez, kutxen grafiko bat eratu nahi badugu, bataz bestekoak, desbiderazioak eta maximo eta minimoak konparatzeko, aukera hori izango dugu. Komando hauek sartzea konplikatuagoa izango da. Horregatik, lehenengo punturaino iristeko sartu behar diren komandoen jarrai puntuak ezagutzea:

histogram(~Sepal.Width|Species,data=iris,strip=strip.custom(bg="lightgrey",par.strip.text=list(col="black",cex=.8, font=3)),main="Distribution of Widths by species", xlab="Height (inches)")

Irudia 16Lattice paketearekin sortutako histograma Rrekin landuta. Iturria: egilea



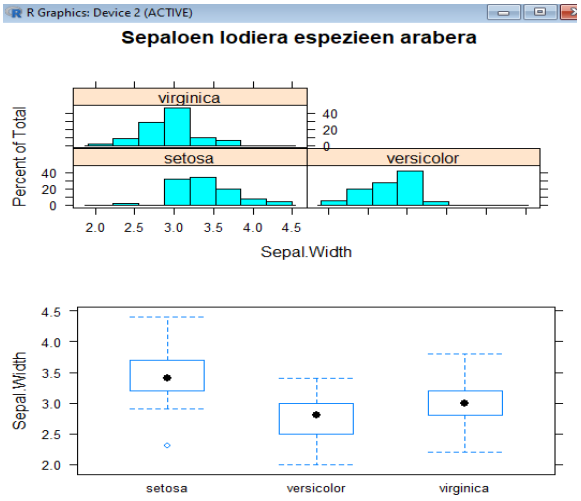
Lehen egindako grafikoan bezela, *Psych* paketean, desberdintasun batzuk nabaritu ahal izango ditugu, espezie ezberdinen artean. Baino ez dira hai nabariak izango, petaloen lodiera bezalakoak. Beraz, esan dezakegu, sepaloaren lodierak ez digula pista haundirik emango, espeziea zein izango den jakiteko.

Beste grafiko batekin partekatzeko, *graph1* eta *graph2* grafikoak definituko ditugu, komandoak sartu baino lehenago:

```
> graph1 <- histogram(~Sepal.Width | Species, data = iris,main = "Sepaloen lodiera espezieen arabera" )

> graph2 <- bwplot(Sepal.Width~Species, data = iris)

> plot(graph1, split = c(1, 1, 1, 2))

> plot(graph2, split = c(1, 2, 1, 2), newpage = FALSE)
```

Emaitza honako hau izango da:

Irudia 17Espeziearen arabera, sepaloen lodiera Rrekin landuta. Iturria: egilea



Ikusi daitekeenez, kutxa diagramaren laguntzarekin, datuak hobeto ulertu daitezke kasu honetan. Sepalo lodiera haundiena duten datuak, *Setosa* espeziekoak izango dira, azterketaren arabera, eta txikienak *Versicolor* espezietakoak. Gainera ikusi daiteke *Versicolor* eta *Virginica* espezieen artean, sepaloaren lodieran desberdintasun oso txikia dagoela. Beraz, bi hauek nabaritzeko faktore hau erabiltzea ez da aproposa izango.

Esan dugun bezala, pakete hau ezinbestekoa izango da, guk deskribatu behar dugun grafikoa era zuzenean sortzeko. Ondoren kasu praktikoa eratzerakoan, pakete hau ezinbestekoa izango da. Bigarren grafikan, guk kutxa diagrama erabili dugun lekuan, edozein grafika barneratu dezakegu. Gure

kasuan kutxa diagrama ipini dugu, informazio gehiena ematen zigun grafika zelako. Beste kasu batzuetan, beste aukera batzuk ipintzea erabilgarriagoa izango da. Kontuan hartu beharko dugu, beti, guk egiten ditugun grafika guztiak, irakurlearen ulermenerako baliagarriak izan behar direla.

4.4.3 GGPLOT2 paketea

R programarekin grafikoak egiten dutenek, pakete honen erabilpena gomendatzen dute. Grafikoak editatzeko eta sortzeko aukera ugari emango dizkigun paketea izango da. Ardatza mota ezberdinak, tema ezberdinak,…Baita ere beste pakete batzuk, pakete honen aurre instalazioa beharko dute, hauekin lan egiteko ondoren *Patchwork* paketearekin ikusiko dugun bezala.

GGPLOT2 paketeak, grafiko ikusgarriagoak egiteko aukera ezberdinak emango dizkigu. Orain, *iris* datu basearekin ez dugu lan egingo. Emaitzak hobeto ikusteko, beste datu base bat aztertuko dugu, pakete honetan bertan dagoena. Datu base honetan, 58788 pelikulari buruzko informazio desberdina agertzen da, zein nota duten, izenburua, publikazio data, gastatutako dirua, generoa…24 aldagai guztira.

Ikusi daiteke, *GGPLOT2* paketeak sortzen duen histograma, zertxobait desberdina dela. Datuak interpretatzeki marra horizontal eta bertikal ezberdinak jarriko ditu, histograma hobeto ulertzeko. Era sinple batean grafiko bat egingo bagenu, horrelako grafika bat geratuko lirateke, komando hauek erabiliz:

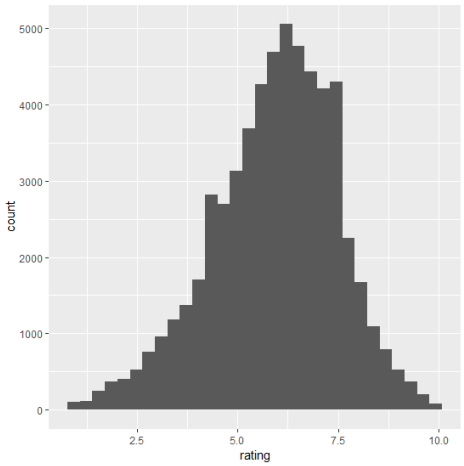
```
> pelikulak<-ggplot(movies, aes(x=rating))

> pelikulak+geom_histogram()
```

Grafiko honetan ikusi daiteke histograma barnean dauden ardatzak asko laguntzen digutela, informazioa ondo prozesatzeko, baina garrantzitsua izango da, tarteak ondo definitzea eta bordeak marraztea tarteen artean. Grafiko txukunago bat bilatzeko, komando hauek jarriko ditugu, irakurleari edo ikerlariari laguntzeko:

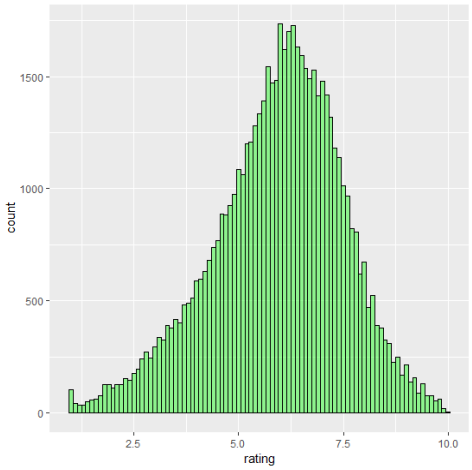
```
> pelikulak2<-pelikulak+geom_histogram(binwidth
= 0.1, col='black', fill='green', alpha=0.4)
```

Hurrengo bi irudiak, guk sortutako bi grafikak izango dira. Lehenengoa, *pelikulak+geom_histogram()* komandoa erabili ondoren, eta bigarrena, koloreak eta bordeak barneratu ondoren.

Irudia 18Movies datu baseko pelikulen notak Rrekin landuta. Iturria: egilea



Irudia 19Movies datu baseko pelikulen notak Rrekin landuta. Iturria: egilea

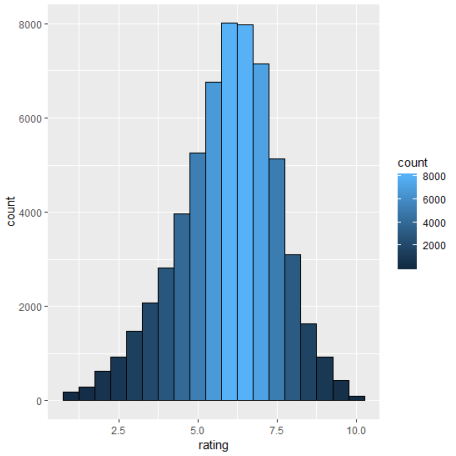


Nahiz eta bi grafikoek informazio berdina izan, ikusi daiteke bigarren grafikoak informazio gehiago duela. Bietan argi nabarituko da distribuzio normal bat jarraitzen dutela, baino bigarrenean, tarteak hobeto ulertuko dira. Gogora dezagun, tarte gehiegi erabiltzea ez dela aproposa izango irakurlearentzat. Agian, hemen ikusi dezakegu zein den nota errepikatuena (dezimal bat kontuan hartuta). Baina informazio gehiegi izango da grafika era egokian ulertu ahal izateko. Beraz, komenigarria izango da, gure grafikan tarte gutxiago erabiltzea. Kalifikazioekin lan egiten dugunean, suspentsoak zeintzuk diren definitzea komentigarria izango da, horregatik, zenbaki borobilekin lan egitea komenigarria izango da. 1eko tarteak agian haundiegiak direnez eta informazioa galdu daitekeenez, 0,5eko tarteak erabiliko ditugu. Gainera, legenda bat jarriko dugu, irakurleak erreferentzia hobeak izan dezan. Komando hauek erabiliko ditugu:

```
> PELIKULAK3<-pelikulak+geom_histogram(binwidth
= 0.5, col='black', fill='green', alpha=0.4)
```

```
>PELIKULAK3+geom_histogram(binwidth=0.5,aes(fill=..count..), col='black')
```

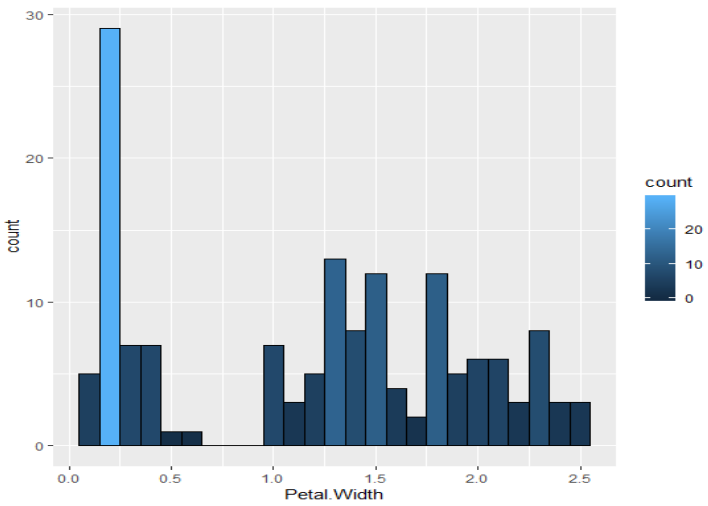
Irudia 20movies datu basetik lortutako histograma, tarteen zabalera egokia erabiliz. Iturria: egilea



Nahiz eta histogramak berak zenbatekoa definitu, koloreak lagundu egingo digu, datu multzoa azkarrago aztertzen. Ikusi daiteke 5etik gorako pelikula gehiago daudela, 5eko puntuazioa baino baxuagoa dutenak baino.

Iris datu basearekin lan egingo bagenu, histograma horrela geratuko lirateke.

Irudia 21Prozesu berdina jarraitu dugu iris datu basearekin. Iturria: egilea



Ikusi dugunez, GGPLOT2 paketeak, grafikoen irudikapenean aurrerapauso handia da. Gainera, beste ekarpen batzuk baditu. Adibidez, google-ekin lan egiterakoan, koordenadak izanda, mapa interaktiboak sortu daitezke, bertan informazioa ulergarriagoa izan dadin, beranduago ikusiko dugun bezala.

4.4.4 Patchwork paketea

Lehen aipatu ditugun beste pakete batzuk, histograma edota grafika ezberdinak batera ikusteko aukera ematen ziguten. *Patchwork* paketeak, aukera hau emango digu, baina konfigurazioa gure erara bideratzeko aukera emango digu, hau da, grafiko bakoitzaren posizioa eta garrantzia zehazteko aukera, ondoren ikusiko dugun bezala:

Garrantzitsua izango da pakete hau deskargatzeko, ez da beste paketeen deskarga bezala eratuko. Lehenengo *devtools* paketea deskargatu beharko dugu, eta ondoren, komando hau erabili beharko dugu:

```
>devtools::install_github("thomasp85/patchwork
")
```

Erabiltzaile partikular batek garatu duen paketea denez era honetan deskargatu beharko dugu. *Github* web orrialdean, 36 milioi erabiltzaileek kode irekiko programekin lan egiten dute eta beraiek garatutako programekin lan egiteko aukera izango dugu. Kasu honetan *Thomasp85* erabiltzaileak garatutako paketea erabiliko dugu. Azpi orrialdean29 eskaintzen den linkean, ikusi daitezke erabiltzaileek garatu dituzten programa ezberdinak.

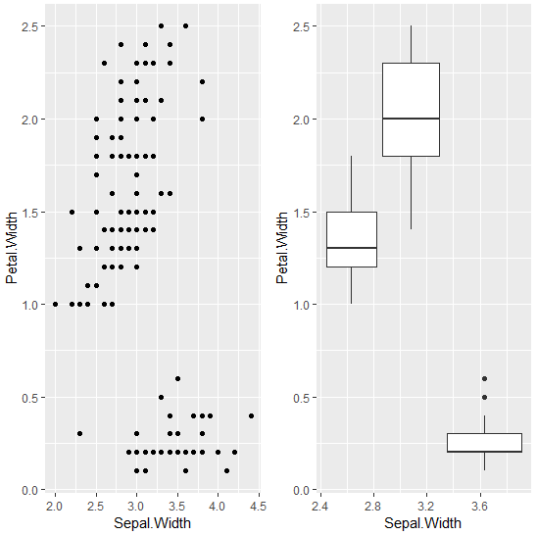
Beraz, pakete honek eskaintzen dituen aukerak ikusteko, erabil dezagun *GGPLOT2* paketearen datu base bat. Beraz, *GGPLOT2* eta *patchwork* paketeak deskargatu beharrak izango ditugu. Hurrengo komandoak erabili beharko ditugu grafiko hau lortzeko:

```
> library(ggplot2)
```

[29] GitHub. (2018) *thomasp85/patchwork*. Github.com. Iturria: .https://github.com/thomasp85/patchwork.

```
> library(patchwork)

>ggplot(iris)+geom_point(aes(Sepal.Width,Petal.Width))

>i1<ggplot(iris)+geom_point(aes(Sepal.Width,Petal.Width))

>ggplot(iris)+geom_boxplot(aes(Sepal.Width,Petal.Width,group= Species))

>i2<ggplot(iris)+geom_boxplot(aes(Sepal.Width,Petal.Width,group= Species))

> i1+i2
```

Irudia 22Patchwork paketearen aukerak aztertzen Rrekin. Iturria: egilea



Aukera honek, datu base baten elementu ezberdinak konparatzeko balio izango digu. Alde batetik, puntu diagrama bat izango dugu, datu guztien distribuzioa nolakoa den ulertzeko, eta bestetik, espezie bakoitzak izango duen distribuzioa ikusiko dugu. Era honetan, argi ikusiko da adibidez, petalo lodiera 0,5ekoa baldin bada, espezie batena izango dela, kasu honetan, lehen aipatu dugun bezela, *Virginica* espeziea izango dela baieztatu ahal izango dugu.

Lehen, grafikak alderatzeko pakete ezberdinak daudela aipatu dugu. Pakete honen ekarpen nabarienetakoa hauxe bera da, grafikoak partekatzeko

aukerak emango dizkigu. Zein da pakete honen abantaila? Erantzuna sinplea izango da. Pakete honek, gure grafikak nahi dugun moduan alderatzeko aukera emango digu, horrela, guk gure modura antolatzeko aukera izango dugu.

Adibidez, *plot_layout* komandoak, bigarren grafikoa azpikaldeko posizioa hartzea bideratuko digu. Bi grafikoen artean distantzia haundiagoa egotea nahi baldin badugu berriz, *plot_spacer* komandoa erabili beharko dugu. Horrela, grafiko bat gure gustora moldatu ahal izango dugu. Pakete honek, bi grafiko baino gehiagorekin lan egiteko aukera emango digu, hauek ere, gure gustura ipiniz. Hurrengo komandoekin, grafiko gehiago sortuko ditugu, *i3* eta *i4* izenekoak (garrantzitsua izango da, gure lanerako, laburdura sinpleak erabiltzea, datu batzuk behin eta berriz tratatu behar badira, bestela komando luzeegiak izango dira erabiltzaile ikasle batentzat). Komandoak hauk izango dira, beraz:

```
>i3<ggplot(iris)+geom_smooth(aes(Petal.Length, Petal.Width))
```

```
> i4<-ggplot(iris) + geom_bar(aes(Petal.Width))
```

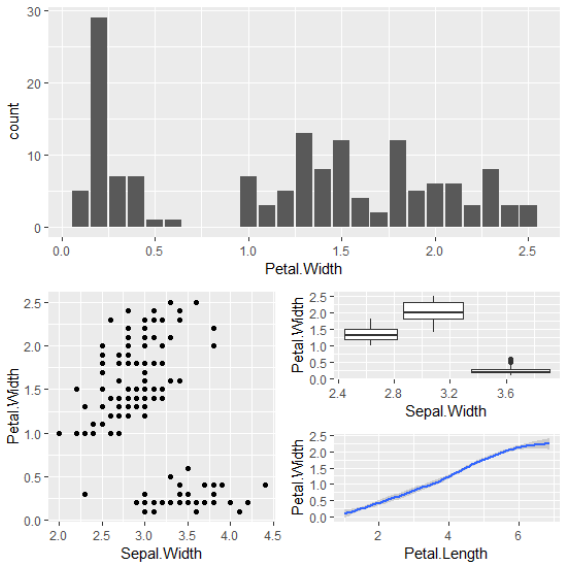
Desberdintasun nabarmena dago bi grafiko hauen artean, baino grafiko gehiago sartu ezkero, aukerak nabarmenagoak izango dira, eta garrantzitsua izango da informazioa era egokienean aurkeztea. Garrantzi haundien duen grafikoa lehenengoa jarri, azpigrafikoak garrantzi baxuagoa izanda,…Ikus dezagun beste bi aginte sartu ezkero nola geratzen den:

```
> p3 <- ggplot(mtcars) + geom_smooth(aes(disp, qsec))
```

```
> p4 <- ggplot(mtcars) + geom_bar(aes(carb))
```

```
> p4+ {p1+{p2+p3 + plot_layout(ncol =1) }}+ plot_layout(ncol = 1)
```

Irudia 23Iris datu basearen aldagaien azterketa, grafika ezberdinak konbinatuz. Iturria: egilea



Ikusi dugun bezala, pakete honek ekarpen haundi bat suposatuko du. Grafikoen posizionamendua askotan oso garrantzitsua izango da, informazioa era egokian transmititzeko. Nahiz eta grafiko gehiago ez barneratu, grafiko gehiago barneratzeko aukera izango dugu, nahi adina grafiko. Beraz, tresnaa egokiak erabili ezkero, pakete honekin gure grafiken optimizazioa bilatzen saiatuko gara. Datuak tratatzean, maiztasun poligonoak eratzeko aukera ezberdinak aztertzea gomendagarria izango da. Aukera hauetako bat, *agricolae* paketea deskargatzea izango da. Bere komandoak lagungarriak izango dira, gure grafikak hobeagoak izateko, ondoren ikusiko dugun bezala.

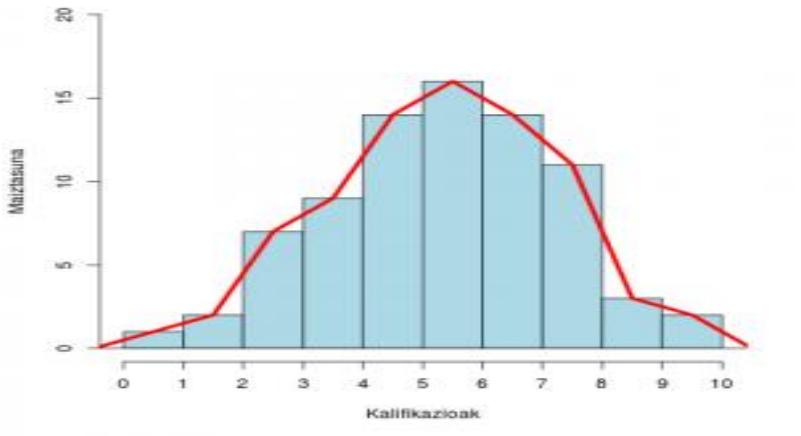
4.4.5 Agricolae paketea

Histogramen eraketan, elementu aipagarri bat aipa dezakegu; maiztasun poligonoa[30]. Frekuentzia polígono hau osatzeko, histogramako barrak, bere tarteen zentroetatik pasatzan den marra batekin elkartuz lortzen den figura izango da. Figura hau lortzeko, Rtik *Agricolae* paketea deskargatu behar izana daukagu lehenago, eta *Polygon.freq* funtzioa erabili behar izango

[30] Maiztasun-poligonoa edo maiztasunen poligonoa aldagai estatistiko kuantitatibo jarraitu baten maiztasun-banaketa adierazten duen grafiko estatistiko bat da, datu gehienak zein tartetan biltzen diren adierazten duena. Histogramatik eratortzen da zuzenean, zutabeko erdipuntuak lotuz, eta hura bezala interpretatzen da. Maiztasun-poligonoa leunduz, maiztasun-kurba delakoa sortzen da. (Gizapedia)

dugu. Honekin, modeloak jarraitzen duen distribuzioa ezagutzeko aukera izango dugu. Ikusi dezagun, maiztasun polígono batek izango duen forma:

Irudia 24Maiztasun poligono batek duen forma Iturria:Gizapedia



Ikusi daitekeenez, informazio ezberdina irudikatzeko gaitasuna emango digu polígono frekuentziak. Alde batetik, maiztasun haundiena duen tartea nabarmenduko da beste guztien artean eta gainera, barneratutako datuen banaketa aztertu ahal izango dugu. Banaketa honekin, azterketa estatistiko ezberdinak egin daitezke, hipotesiak burutu eta kontrastatu,..Adibidez, irudian ikusten den banaketa, *banaketa normal* bat jarraitzen duela argi ikusiko dugu, gehien errepikatzen diren datuak erdikoak direlako eta zenbat eta batezbestekotik hurrunago egon, zenbat eta maiztasun baxuagoarekin errepikatuko dira datuak. Hemen ikusi ditzazkegu frekuentzia poligonoen eraketak eskaini ditzakegun abantaila batzuk:

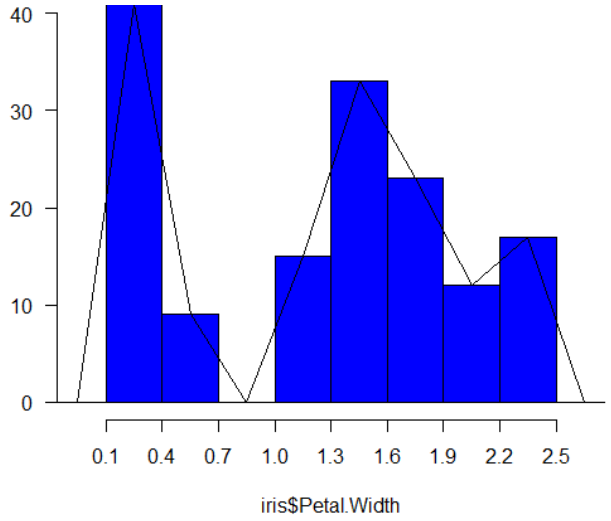
- Parametro ezberdinak erabili daitezke grafika berean ondorio ezberdinak ateratzeko.
- Datu ezberdinak grafika berdin batean biltzen ditugu. Era honetan, datu multzoen grafikoak espazio gutxiago okupatuko dute. Oso baliagarria izango zaigu, aurkezpenetarako adibidez.
- Tresna ezberdinak erabili daitezke, *Paretoren diagrama*[31] adibidez, distribuzioa ezagutzeko.

[31] https://support.minitab.com/es-mx/minitab/18/help-and-how-to/quality-and-process-improvement/quality-tools/supporting-topics/pareto-chart-basics/#what-is-a-pareto-chart link honetan eskuragarri izango dugu, Paretoren Diagramari buruzko informazio gehiago.

Beraz, R programa informatikoarekin era zuzenean lan egiteko, *agricolae* paketea deskargatu eta instalatu beharko dugu. Gogora dezagun, *Iris* datu basearekin ari garela lanean. Beraz, frekuentzia poligonoa eratzeko, hurrengo komando hauek erabili beharko ditugu, emaitza, honako hau izanez:

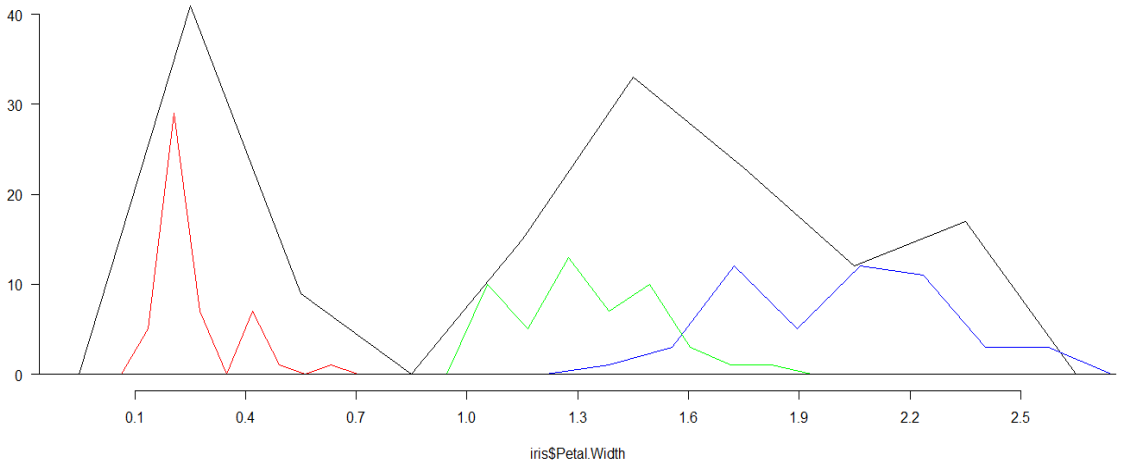
```
> library(agricolae)

>maiztpol<Graph.freq(iris$Petal.Width,col="blue")

> polygon.freq(maiztpol,col="black",ltd=4)
```

Irudia 25Maiztasun poligonoa Iris paketean Rrekin landuta. Iturria: egilea



Ikusi dugun bezela, maiztasun poligonoa lagungarri izango zaigu, baino grafiko berdinean, bi maiztasun poligono baino gehiago barneratu ahal izateak, gehiago lagunduko digu, 3 espezieak konparatu ahal izateko. Komenigarria izango da, 3 espezieen arteko konparaketa egiteko, 3 espezieen maiztasun poligonoak lortzea. Horretarako, eta errazago ikusi ahal izateko, barrak ez egotea komenigarria izango da, informazioa era egokian eman izan dadin. Kolore bakoitza espezie bat suposatuko du, eta beltzez dagoen marra, maiztasun absolutua izango da. Horrela, irakurleak informazio ezberdina lortu ahal izango du, bai espeziearen barruan dauden ezaugarrien artean, eta baita ere orokorrean atera daitezkeen ondorioak, distribuzioa, batazbestekoa,…

Irudia 26Iris datu baseko maiztasun poligonoa eta espezie bakoitzeko maiztasun poligonoa Rrekin landuta. Iturria: egilea



Ikusi dugunez, *Agricolae* paketeak, laguntza ugari eskaintzen dizkigu. Maiztasun poligonoekin marrak egiteko aukera, grafika ezberdin asko konparatzeko aukerak emango dizkigu. Ez da berdina izango bi histograma bata bestearekin konparatzea, bata bestearen gainean jarrita. Bertan informazio asko galduko da, eta grafikoetan saiatu behar garen gauzetako bat, informaziorik ez galtzea izango da.

4.5 HISTOGRAMAREN EDIZIOA

4.5.1 KOLOREAK

Grafiko baten helburua, estatistika deskribatzailearena izango da. Lehenago aipatu dugun bezala, estatistika deskribatzaileak bilatuko duena, ikertzailearen edo irakurlariaren datuen interpretazio azkarragoa eta zahatzagoa izango da. R programak, histograma eratzeko komando estandarra sartu ezkero, txuri beltzean eratutako histograma sortuko digu. *Softwareak*, edozein koloretan editatzeko aukera emando digu. Ikusiko dugunez, 3.5.2 bertsioan barneratuta egongo dira (*lattice* paketearekin batera) kolore basiko batzuk.

Puntu honetan ikusiko duguna, zein aukera izango ditugun grafikoen koloreak editatzeko izango dira. Funtzio batzuk, bakarrik grafiko txukunago bat burutzeko balioko dute, baino beste pakete batzuk, erabilpen zehatz bat izango dute. Adibidez, ondoren ikusiko dugun aukeretako batek, daltonismo

problemak dituzten pertsonei zuzenduta dago. Pakete hau, ez dator R sisteman integratuta, baino interneten aurkitu ditzazkegun paketeetako bat izango da. Ikus dezagun hasieran, nola aldatu daitezkeen grafikoetako koloreak:

Grafiko bat koloreztatzeko, *col=”kolorearen izena”* komandoa ipini behar dugu. Beraz, *hist(altuera,col=”red”)* komandoa erabili ezkero (altuera, lehen aipatutako datuak erabiliko ditugu. R *softwarean* manualki sartu direnak). Komando hau, oso generikoa da, edozein grafiko motatan erabili daitekeena. Horregatik, grafiko guztietan ez du efektu berdina izango, eta agian komando espezifikoagoak erabiltzea komenigarria izango lirateke.

Orrialdearen oinean[32] *Columbiako unibertsitateak* burutu zuen kolore guztien bilketa ikusi dezakegu. Ikusi daitekeenez, kolore asko barneratzen ditu pakete basikoak. Kolore hauek, grafikoa gure erara moldatzeko aukerak emango dizkigute. Demagun, ardatzen kolorea urdina izatea nahi dugula, etiketen kolorea verdea, izenburuaren kolorea gorria eta azpitituluaren kolorea arroxa izatea nahi dugula. Aipatu ditugun aldaketak burutzeko, honako komandoak erabili beharko genituzke:

```
>Col.axis=”blue”
>Col.lab=”green”
>Col.main=”red”
>Col.sub=”pink”
```

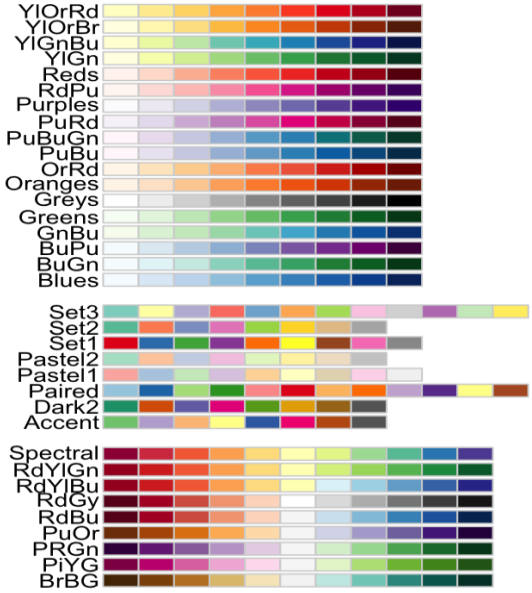
Denbora asko galdu ez nahi badugu, programak berak tema desberdinak proposatuko ditu gure grafikarentzat. Adibidez, *col=”rainbow”* komandoak ostadarraren koloreak erabiliko ditu gure grafikoan. Erabiltzaileek pakete ezberdinak igo dituzte *softwarera*, hauekin kolore modelo ezberdinak erabili daitezke. Hauen artean aipagarriena, *RColorBrewer* paketea izango da.

Pakete honek, kolore sekuentzia ezberdinak eskainiko dizkigu, mota eta tamaina ezberdinetakoak. Oso komenigarria izango da, beraz, pakete hau

[32] Columbiako Unibertsitatea. (2014). *Colors in R*. Iturria: Stat.columbia.edu http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf.

erabiltzea grafikoaren diseinuan denbora gutxi baldin badugu. Hemen ikusi ditzazkegu pakete honek dauzkan aukera batzuk:

Irudia 27RColorBrewer paketeak izango dituen kolore transizio ezberdinak
Iturria: Columbiako Unibertsitatea



Estatistika deskribatzailearen zientzia garatu nahiean, 2015 urtean *Thomas Lumleyek* daltonismoa duten pertsonentzat adaptatutako koloreak moldatuko ditu, eta *Dichromat* paketean bildu. Pakete hau, nahiz eta gure grafikoan aldaketarik ez sortu, oso garrantzitsua izango da, grafika daltonismoa duten pertsonei aurkeztu behar badiegu. Beraz, komenigarria izango zaigu pakete hau ere deskargatzea. Ikus dezagun bi grafikoen artean egongo dan diferentzia *Dichromat* paketea arabili ezkero:

Imagina dezagun grafika ikusi behar duen pertsonak *deuteranopia*[33] duela. Grafikoaren helburua, lehenago aipatu dugun bezala, irakurleak jasotzen duen informazioa ahal den eta ulermen haundiagoa izatea izango da. Horretarako, *dichromat*[34] paketea deskargatu beharko dugu. Hurrengo irudian

[33] Daltonismoa izatean egon daitezkeen aukeretako bat da. Deuteranopia dutenek, kolore berdea ikusteko problemak izaten dituzte. Horregatik, kolore berdea grafikoetatik ebatzi beharko dugu.
[34] Lumley, T. (2015). *Color Schemes for Dichromats*. Iturria: CRAN

argi ikusi ahal izango dugu, koloreen adaptazioa nolakoa izango den, eredu berria ikusita:

Irudia 28Dichromat paketearen kolore eskala Iturria: CRAN

Gainera, behin pakete basikoak ezagutu ditugula grafikoen koloreen edizioan, aipagarria izango da, *Munse*ll paketearen ekarpena. Batzuetan zaila izango denez koloreekin asmatzea, eta urdin bat bilatu nahi badugu adibidez, guk bereziki nahi dugun kolore urdin hori bilatzea ez da eraza izango. Askotan, koloreek, talde bate do kolore bat erabiltzen duen edozein kolektibo errepresentatzeko erabiltzen da funtzio hau. R pakete hau deskargatu behar izango da. Nahi dugun urdina aukeratzeko, adibidez, aukera ezberdinak izango ditugu. Hemen ikusi ditzazkegu kolore urdinak izan ditzazkeen kasu batzuk. Hau edozein koloreekin egin ahal izango dugu:

Irudia 29Munsell paketearen aukera ezberdinak urdin kolorearekin. Iturria: CRAN

Honetaz aparte, fondoaren kolorea aldatzeko aukera izango dugu, histogramaren enkuadraketa, programak berak ere, histograma tema batzuk ekarriko ditu berarekin *GGPLOT2* paketeak. Oso gomendagarriak izango dira programazio ezagutza baxua duten erabiltzaileek grafiko aurkezgarriak eratzeko. Beti bezala, gomendagarria izango da manualki erabakitzea edozein elementuren kolore edo posizioa, baino denbora gutxi dagoen kasuetarako

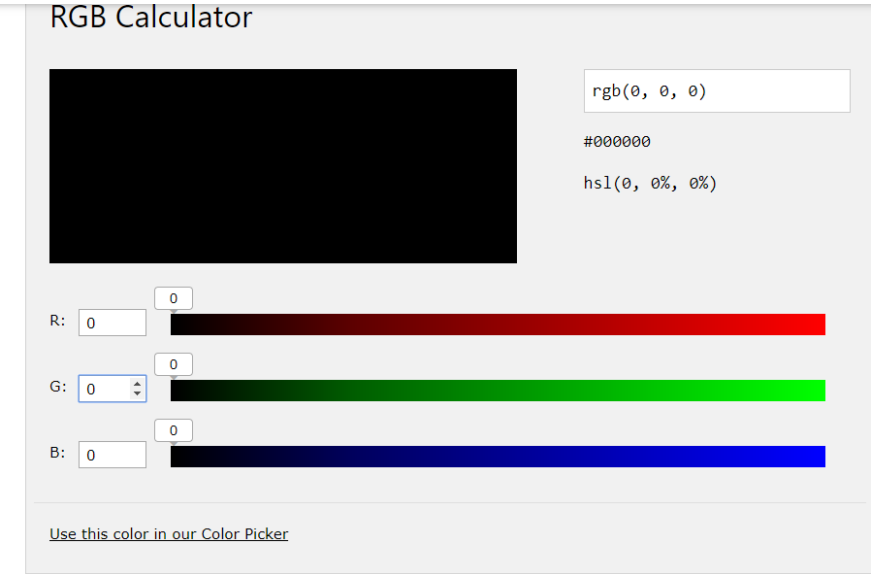
https://cran.r-project.org/web/packages/dichromat/dichromat.pdf

aproposak izango dira. Koloreak ez dira alda daitezkeen grafikoetako elementu bat. Askotan aipatu dugun bezala, R *softwareak, software* librearen izena hartzen du, grafikoen edizioan aukera ia infinitoak eskaintzen dizkigulako. Koloreak, agian, grafikoen errepresentazioan egon daitekeen alderdi garrantzitsuenetakoa izan daiteke, baina beste elementu batzuk (letra tipoa, leyendaren eraketa,…) aldatzeko edo ipintzeko aukerak izango ditugu. Ikus dezagun beraz, lerroak editatzeko izango ditugun aukerak. Aipagarria izango da *Paul Murielek* egindako lana, grafikoen edizioan sortutako gida (*R Graphics*), gure lanaren oinarri bezela hartu dugula. Koloreekin aukera gehiago jakin nahi ezkero, liburu hau osagarri bezela gomendagarria izango da.

Koloreekin bukatzeko, erabiltzaileak askotan, kolore espezifiko batzuk erabiltzeko ohitura edo beharra izan dezake, eta kolore hau lortzea zaila suertatu daiteke. Programaren garapenean lan egiten duten erabiltzaileek, nola ez, problema hau suertatzeko aukera emango digute, oinean ikus daitekeen *web orrian*[35]. Hiru kolore primarioetatik edozein kolore atera daiteke, eta web horriak berak hiru kolore primarioak elkartzeko aukera emango digu. Aukeraketa hau egin ostean, kolorearen kodea zein izango den agertuko zaigu, eta bertan testu beltza eta testu txuria nola geratzen diren ikusgarri izango dugu. Puntu hau, batzuentzat oso nabarmena suerta ez daiteke, baina grafiko batzuk eratzean, kolore bakoitzak elementu bat errepresentatzen badu, beharrezkoa izango da errepresentazio zehatza izatea. Demagun grafiko bat eratu dugula telefonia marka erabilienekin, eta *Vodafone* markaren kolore gorri bizia imitatu nahi dugula. Web orri honek kolorea kopiatzen lagunduko digu. Ikus dezagun erabiltzaile batek izango duena ikusgarri:

[35] https://www.w3schools.com/colors/colors_picker.asp web orrian koloreak aukeratzeko sistema bat izango dugu. Bertan, edozein kolore sortzeko aukera izango dugu.

Irudia 30Koloreak sortzeko tresna. Iturria: RPubs



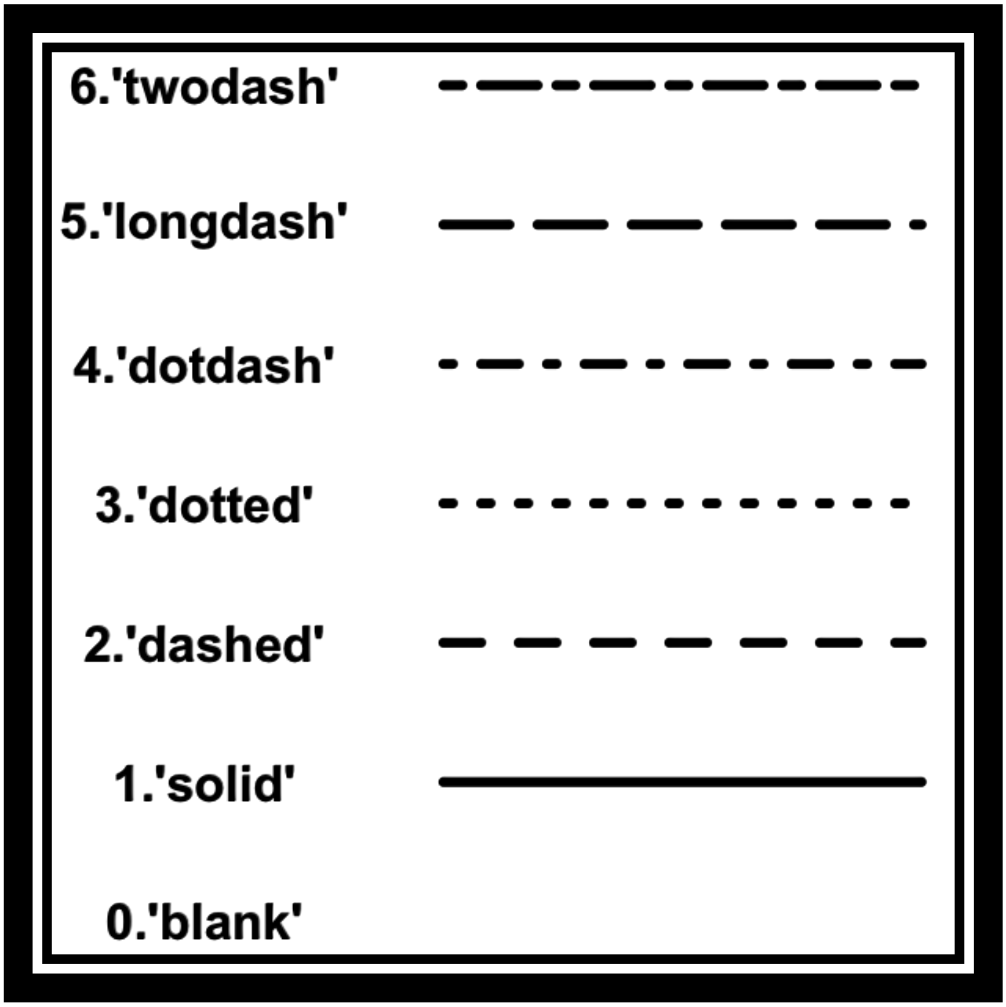
Ikusi dugun bezala, R programak aukera desberdin asko emango dizkigu, gure grafikoa kolore egokiekin erabili ahal izateko. Beraz, programa honek, koloreen kasuan mugarik gabekoa izango dela esan ahal izango da, beste programa batzuk ez bezala (*Oracle*, adibidez). Beraz, erabiltzailearen imaginazioaren pean egongo da, grafikoaren diseinua egokia izatea.

4.5.1.1 Hrbrthemes paketea

4.5.2 Lerroak

Lerroak, ez dira koloreak bezain ikusgarriak, eta ez da hainbeste garatu lehen aztertutako elementua bezala. Hemen, programak bi aukera eskeiniko dizkigu *lty* eta *lwd* komandoak izango direnak. Lehenengoak lerro motadefinituko duen komandoa izango da. Adibidez, *lty=”dotted”* agindua *softwarean* idatzi ezkero, grafikoen lerroak, puntuz osatutako marrak izango dira. Komandoak azkarrago idatzi nahi ezkero, zenbaki bidez predefinituta izango ditugu, agindua azkarrago idazteko. Rn garapenean lan egiten duten *Sthda* webgunean, lerro moten aukeren aginduak eta beren zenbakiak errazten dizkigute:

Irudia 31Rk eskaintzen dituen lerro motak Iturria: CRAN



Honetaz aparte, *lwd=1* komandoak pixel baten zabalera duen marra bateko lerroa izateko agindua izango da. Lerroak nabarmenak izatea nahi badugu, lwd haundiagoa idatzi beharko dugu kontsolan.

Nahiz eta oso aldaketa ikusgarriak ez izan, gure grafikak hobetzeko erabilgarria izango zaigu kasu batzuetan. Marren edo lerroen lodiera haunditu ezkero, histogramen mugak nabarmendu egingo ditugu, desberdintasunak nabariagoak izan daitezen.

5 ONDORIOAK

Ikusi dugun bezala, ez da zaila izango edozein programa informatiko erabiltzea, grafikoak eratzeko. Problemak, grafikoaren eraketan eta programaren erabilera egokian etorriko dira. Horregatik, garrantzitsua izango da, grafiko bat sortu baino lehenago, grafiko hori nola egin behar dugun jakitea, hau da, tarteak era egokian kartzea, ardatzak zeintzuk izango diren definitzea, tarte erregular edo irregularrak…Ez dago metodo espezifiko bat esateko grafika bat ondo dagoela baieztatzeko. Batzuetan, grafika batek informazio bat erakustea bilatuko dugu, eta beste batzuetan, beste informazio bat erakusteko aukera izan dezakegu, informazio berdina erabiliz. Laguntza moduan, matematikariek tarteak definitzeko aukera ezberdinak proposatzen dizkigute, lehen ikusi dugun bezela, gure grafikak efizienteagoak izateko. Baina kontuan hartu behar dugu, estatistika deskribatzailean pertsona bat edo gehiagorentzat prestatzen direla datuak, eta batzuetan aldaketaren bat egin behar dela, matematikak proposatzen dizkigun ereduetarako. Gogora dezagun, abiadurekin egindako grafika, nola nahiz eta matematikak tarte zabalera bat eman, gure kasuan 5 km/htara aldatu dugu, irakurleak erreferentzia argiagoak izango ditu, abiadura topea zenbat jende pasatzen duen ikusteko, adibidez. Kasu hauek, askotan errepikatu daitezke. Matematika, erreminta bat izango da, erabili beharko duguna. Hau erabili ondoren, ikusi beharko dugu irakurle edo ikertzaileak informazioa era egokian jasoko duen, matematikak proposatzen digun eredua jarraituz. Horretarako, garrantzitsua izango da, gure grafikekin zein mezu bidali nahi dugun jakitea.

Behin grafikoa nola egin nahi dugun irudikatu dugula, programa informatiko baten beharra izango dugu. Hemen erronka bat izango dugu, programaren aukeraketan. Lan honetan, programa ezberdinak mintzatu ditugu eta hauen artean R programa aipatu dugu, programa aproposena bezala. Nahiz eta programa errazagoak egon (*Excelen* bertan grafikoak egiteko aukera ezberdinak egon daitezke), garrantzitsua izango da programa batek aukera asko eskaintzea. Ikusi ditugun programak, batzuk ordaintzekoak zirenez, gastu ekonomiko haundia suposatu dezakete empresa batentzat, batez ere txiki eta ertaientzat. Beraz, 0ko kostu ekonomikoa duen eta grafikoak garatzeko aukera nabariak eskaintzen dituzten programen artean R aukeratu dugu.

Software honek dituen abantailaz aparte, bere konpetentzia diren programekin bere desabantailak ere izango ditu. Alde batetik, komandoekin lan egiten duen programa denez, garrantzitsua izango da programazio jakintza minimo batzuk edukitzea. Horretarako, Rren erabiltzaileek *Rcmdr* paketea garatu zuten. Pakete honek, menu bat sortuko du, zein komando basikoak garatzeko aukera emango digu, baina grafiko garatuagoak nahi egin ezkero, erronka berdina izango dugu. Erabiltzaileak *S hizkuntza informatikoaren* ezagutza izan beharko du programarekin jarraitzeko. Gainera, erabiltzaile basiko batentzat ez da erraza izango programa ulertzea eta komandoak era egokian exekutatzea. Zorionez, lanean aipatu dugun bezala, sistema libre bat denez, erabiltzaileek sistema hau garatzeko eta pakete ezberdinak sortzeko aukerak izango dituzte. Beraz, erabiltzaileen kolaborazioa oso lagungarri izango da, programaren garapenerako. Horretaz aparte, lehen aipatu dugun bezala, programazio nozio batzuk izatea beharrezkoa izango da. Aipagarria da programa honek duen dokumentazio ofizial guztia. Pakete bakoitzak bere gidaliburua izango du, *PDF formatoan* guztiz eskuragarri. Tutorial bakoitzean, pakete horren komando guztien funtzioak azalduko dira, adibideekin, erabiltzaileak erraz uler dezan, eta bere lanean aplikatzeko gai izan dezan. Beraz, nahiz eta sistema konplikatua izan, programa landu nahi ezkero, interneten informazio ofizial asko aurkitu dezakegu, zein gure lanean asko lagunduko digun eta programatzen ikasiko dugun, pixka bat, nire kasua izan den bezala. Lana hastean, *S hizkuntza informatikoari* buruzko ezagutza xumea zen, eta orain esan dezaket, gidaliburuak begiratu ondoren eta *softwarearekin* lan egin ondoren, programarekin lan egiten ikasi detela. Beraz, iruditzen dena baino denbora laburreagoan ikasten dela ikusi daiteke.

Egindako guztia ikusi ondoren, esan dezakegu R programak, datu asko dauzkagunerako aproposa dela. Askotan, datu gutxi ditugunean (5 edo 10 datu) errazagoa izango da ondorio batzuk ateratzea datu horien artean, edota datuak sisteman barneratzea eskuz. Programa honek, milioika datu eskuragarri uzten dizkigu gure paketeetan, eta gainera, dokumentu ezberdinetatik irakurtzeko gaitasuna izango du. Aipagarri izango da, *.csv* formatoa irakurtzeko gai izatea. Lan honetan ikusi dugu, datu base ezberdinak interneten bilatzean, plataforma

askok formato hau erabiltzen dutela datuak errazteko. Beraz, Rk datuak era honetan irakurtzeko gai izatea garrantzitsua izango da.

Beraz, lan honetan histograma bat zer den eta nola eratzen den aztertu dugu. Tarteak definitzeko formula ezberdinak daudela ikusi dugu, eta zein erabili behar den zein daturekin aztertu dugu. Baita ere, histograma bat irudikatzean, konturatu gara irakurlearentzat zuzendutako informazioa, irakurlearentzat prest egon behar duela, hau da, eman nahi dugun mezua, argi irudikatu behar dugu. Horretarako, pakete ezberdinetaz laguntza hartzeko aukera izango dugu. Ikusi dugun bezala, Rk emango dizkigun gaitasunak mugaezinak izango dira. Ondo bilatu ezkero, pakete ezberdinak aurkituko ditugu, zein guk bilatzen dugunerako lagunduko digute. Gure zerbitzurako aproposa den programa izango dela esango dugu.

6 BIBLIOGRAFIA

- Barberá, P. (2017). Access to Facebook API via R. Rfacebook paketearen gidaliburua:

  https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf

- COM UPM, & Santander. (2015). Taller de aplicaciones en R https://www.youtube.com/watch?v=-m6XcgfUKRQ
- D.W. Scott. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.

  https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.103

- De Mendiburu, F. (2020). Statistical Procedures for Agricultural Research. Agricolae paketearen erabilera gidaliburua:

  https://cran.r-project.org/web/packages/agricolae/agricolae.pdf

- Deepayan, S. (2020). Trellis Graphics for R. Lattice paketearen gidaliburua: https://cran.r-project.org/web/packages/lattice/lattice.pdf
- EUSTAT. (2020). *Datos estadísticos de la C.A. de Euskadi*. Eustat.eus. http://www.eustat.eus/indice.html
- EUSTAT. (2018). *PANORAMA DEMOGRAFICO 2018* (p. 10). Eustat.eus. https://www.eustat.eus/elementos/ele0015200/Panorama_demografico_2018/inf0015282_c.pdf
- Freedman, D. eta Diaconis, P. (Abendua, 1981). *"On the histogram as a density estimator: L2 theory"*. Alemania: Springer Verlag. DOI 57 (4): 453–476.
- INE. (2020). Dos siglos de gráficos estadísticos: 1750 - 1950 / 1801 - 1850 / William Playfair (1759-1823). Ine.es. https://www.ine.es/expo_graficos2010/expogra_autor2.htm
- Instituto Nacional de Estadistica. (2015). *Tipos de gráficos*. Ine.es https://www.ine.es/explica/docs/pasos_tipos_graficos.pdf
- Kelmansky, D. (2006) Departamento de Matemática de la facultad de ciencias exactas de Buenos Aires http://www.dm.uba.ar/materias/analisis_de_datos/2006/1/teoricas/Teor2a.pdf
- Kooi R (1980). The Optimization of Queries in Relational Databases. PhD Thesis: Case Western Reserve University

- Lin Pedersen, T. (2019). The Composer of Plots. Patchwork paketearen gidaliburua:
  https://cran.r-project.org/web/packages/patchwork/patchwork.pdf
- Lumley, T. (2015). Color Schemes for Dichromats. Dichromat paketearen erabilera gidaliburua:
  https://cran.r-project.org/web/packages/dichromat/dichromat.pdf
- Mathworks akademia. (2017). Es.mathworks.com
  https://es.mathworks.com/content/dam/mathworks/tagteam/Objects/9/90593_92300v01_TAHFactSheet_ES.pdf.
- Newirth, E. (2014). ColorBrewer Palettes. RColorBrewer paketearen paleta eskuragarri:
  https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf
- Playfaire, W. (1786). The Commercial and Political Atlas. London: Cambridge University
- Raymond, E.S.. (1999) The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary, Cambridge (MA),
- Reed, D. (2014). Data smart: Using data science to transform information into insight. *Journal Of Direct, Data And Digital Marketing Practice*, *15*(4). https://doi.org/10.1057/dddmp.2014.33
- Ruiz Soler, M. eta López González, E. (2009). El entorno estadístico R: ventajas de su uso en la docencia y la investigación. Granada. Granadako unibertsitatea.
  https://revistadepedagogia.org/wp-content/uploads/2009/01/243-03.pdf
  TIOBE Programming Community Index. (2017). *R | TIOBE - The Software Quality Company*. Tiobe.com
  https://www.tiobe.com/tiobe-index/r/
- Wickham, H. (2010). GGPLOT2: elegant graphics for data analysis. Springer. New York. The comprehensive R Archieve Network. R Cran. Rren deskarga linka:
  https://cran.r-project.org
- William, R. (2020). Procedures for Psychological, Psychometric, and Personality Research. Psych paketearen erabilera gidaliburua:
  https://cran.r-project.org/web/packages/psych/psych.pdf