Metode grupiranja podataka

Terzić, Deni

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Deni Terzić

METODE GRUPIRANJA PODATAKA

ZAVRŠNI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Deni Terzić

Matični broj: 44996/16–R

Studij: Informacijski sustavi

METODE GRUPIRANJA PODATAKA

ZAVRŠNI RAD

Mentor/Mentorica:

Prof. dr. sc. Jasminka Dobša

Varaždin, lipanj 2020.

Deni Terzić

Izjava o izvornosti

Izjavljujem da je moj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

U ovom radu govori se raznim metodama za grupiranje podataka. Rad će dati uvid kako u teorijski tako i u praktični dio metoda za obradu i grupiranje podataka. Opisano je kako neki od algoritama funkcioniraju isto tako osim riječima metode su prikazane i grafički radi lakšeg poimanja istih. U radu se navode i stvarni primjeri kako bi se lakše prikazalo djelovanje pojedinih metoda u praksi. Napomenuta je važnost korištenja grupiranja podataka u svakodnevnoj ljudskoj djelatnosti kako bi se došlo do novih spoznaja i otkrića. Na kraju rada prikazano je kako je moguće algoritam k-središnjih vrijednosti primijeniti u programu R. Sam primjer u programu provodi se na stvarnim podacima iz baze podataka preuzete s interneta.

Ključne riječi: metode; grupiranje; metoda k-središnjih vrijednosti; analiza; povezivanje; hijerarhijsko grupiranje; nehijerarhijsko grupiranje

Sadržaj

1. Uvod

Razvojem tehnologija došlo je do povećanja broja informacija koje se svakodnevno razmjenjuju putem različitih medija. Navede informacije potrebno je nekako pohraniti te grupirati u određene grupacije. Kako bi se što bolje grupirale informacije postoji više metoda kako bi iste smjestili u odgovarajuće grupe. Budući da se većina informacija može opisati kroz nekoliko obilježja potrebno je dobro analizirati grupe podataka (Institut Ruđer Bošković, 2020.).

Analiza grupa jest metoda multivarijatne analize koja svoju široku primjenu pronalazi u različitim znanostima poput medicine, informatike, različitih društvenih znanosti, i dr. Sama metoda se sastoji od toga da se objekti koji imaju neka slična i/ili ista obilježja grupiraju u istu grupu. Dok s druge strance objekti koji nemaju zajedničke poveznice odnosno različiti su, takvi objekti se grupiraju u različite grupe. Budući da je navedena metoda multivarijatna sami objekti analize promatraju se kroz više varijabli te se tako pokušava odrediti u koju grupu ih je moguće svrstati.

Kako je već napomenuto analiza grupa jest vrlo značajna u različitim područjima ljudske djelatnosti. Tako primjerice u medicini laboranti kod pronalaska novog virusa izoliraju njegov DNK/RNK te po redoslijedu nukleotida ga svrstaju u određenu grupu u kojoj već postoji neki slični virusi. Samim time kad uvrste novi virus u neku od pripadajućih grupa mogu mnogo lakše odrediti grupu cjepiva koji će pomoći u suzbijanju istog.

S obzirom na to kako je analiza podataka jedna od vodećih metoda u svijetu kojom se dolazi do nekih novih spoznaja cilj ovog završnog rada je opisati i objasniti teorijski dio iste te pobliže opisati primjenu na stvarnom primjeru. Kroz ovaj rad proći ćemo kroz više metoda analize podataka kao što je metoda k-središnjih vrijednosti. Isto tako prikazat ću kako prikazati sve obrađene metode u programu R na konkretnom skupu podataka.

2. Metode i tehnike rada

U ovom radu korišteno je nekoliko metoda. Prva metoda izrade rada jest proučavanje literature. Neke izvori literature su prikupljeni tako što su dobiveni od strane mentorice. Dok su drugi pronađeni u gradskoj i fakultetskoj knjižnici te na internetu. Nakon što su pronađeni svi potrebni izvori potrebno je proučiti sve izvore kako bi se izvuklo iz svakoga baš ono što je potrebno za izradu ovog rada. Sam proces proučavanja literature iziskuje od autora izuzetnu koncentraciju te ozbiljan pristup proučavanju iste.

Nakon što se dobro razradi teoretski dio samog rada kreće se u praktični dio rada. Praktični dio u ovom radu jest obrada podataka putem računalnog programa R. Sam računalni program vrlo je jednostavno koncipiran. Kada se otvori program otvara se prozor s konzolom u koju se upisuju naredbe te program sam odrađuje ostalo. Prije same upotrebe programa potrebno se dobro upoznati s razvojnim okruženjem u kojem se nalazi autor. Metoda kako se najbolje upoznati s računalnim program jest taj da se pronađe određena specifikacija za program te uz proučavanje iste paralelno praktično isprobavaju naredbe u programu.

Nakon što se autor upozna s programom potrebno je pronaći podatke nad kojima se vrše metode grupiranja podataka. Kako bi se pronašli podatci autor pronalazi bazu na internetu te istu učitava u program te nakon toga upisuje naredbe koje pokreću određene algoritme. Nakon što program napravi svoj dio odnosno izvrši algoritme zadane od strane autora sam autor kreće u analizu dobivenih podataka. Analizom dobivenih rezultata te njihovim bilježenjem ovaj rad završava.

3. Analiza grupa

Sama analiza grupa jest statistička metoda koja određuje relativno homogene grupe objekata. Primjenu pronalazi u vrlo širokom aspektu znanosti kako bi se utvrdila međusobna povezanost ili različitost objekata same analize. Analiza grupa za svoj konačan cilj ima skupiti objekte u klastere po nekom od promatranih svojstva odnosno varijable analize. Riječ klaster dolazi od engleski riječi *cluster* koja po svojoj definiciju opisuje grupu stvari koje su istog tipa. Tako da i iz naziva može se iščitati kako se u ovoj metodi pokušava grupirati objekte po nekoj zajedničkoj varijabli (Ekonomski fakultet Subotica, predavanje, 2015.).

Tokom analize podataka objekti promatrani u analizi pokušavaju se svrstati u grupu kako po nekoj zajedničkoj varijabli. Tim postupkom dobivaju se homogene grupe ili klasteri koji se razlikuju upravo po toj promatranoj varijabli. Drugim riječima ovakav postupak razlikuje objekte po nekim određenim svojstvima iako se možda oni podudaraju u nekim drugim svojstvima koja se ne promatraju. Ovakvom metodom dolazi do formiranja različitih grupa. Kao krajnji cilj analizom grupa pokušava se kreirati grupe u kojima su objekti što sličniji. Budući da je grupiranje podataka vrlo bitna tehnika kojom se mogu otkriti i razriješiti neki bitni problemi uvrlo kratkom vremenu razvili su se i mnogi algoritmi koji svrstavaju objekte u pripadajuće grupe. Iako svi ti algoritmi u svojoj srži rješavaju isti problem oni su vrlo različiti i ne tako rijetko daju potpuno različita rješenja. Budući da postoji veliki broj algoritama za rješavanje problem nije moguće odrediti koji je najbolji od svih njih. U nekim situacijama jedan algoritam će dati bolje rješenje od drugoga dok će za neki drugi problem pak taj koji je u prvom slučaju bio lošiji biti bolji od algoritma koji je u prvom slučaju bio bolji (Nachtway et al., 2009.).

Analiza grupa jest skupni naziv za cijeli skup multivarijatnih metoda koje kao svoj glavni zadatak pokušavaju svrstati objekte u grupe po njihovim osobinama. Svim tim metodama zajedničko jest da se ne gleda samo jedna varijabla već više njih istovremeno. Najčešće se kao predmet promatranja u analizi grupa uzimaju osobine nekog objekte koje se kasnije percipiraju kao varijable kako bi se što lakše uspoređivali različiti objekti. Same varijable koje ulaze u analizu definira sam istraživač odnosno osoba koja provodi analizu grupa. To svojstvo da istraživač sam određuje koje će varijable promatrati razlikuje analizu grupa od ostalih multivarijantnih statističkih metoda gdje se varijable procjenjuju empirijski (Hair et al., 2010.).

Kao glavni cilj analiza grupa zadaje si grupiranje objekata i definiranje grupa unutar neke skupine objekata. Kod kretanja u analizu samom istraživaču nije poznati koliko će biti konačan broj grupa unutar skupine, niti koji objekt će pripasti kojoj grupi. Kod formiranja grupa potrebno je paziti da se grupe formiraju tako da se u pojedinoj grupi nalaze objekti koji sadrže samo slična svojstva odnosno ne razlikuju se previše. Kako bi se pravilno oformile grupe

unutar pojedine skupine kod analize grupa govorimo o tri pitanja na koja svaka analiza mora pravilno odgovoriti. Ta pitanja su sljedeća: kako mjeriti sličnost između objekata, kako formirati grupe te kako utvrditi konačan broj grupa. Osim ta tri najvažnija pitanja na koja analiza grupa traži odgovore postoji i pravilan redoslijed koraka koji je potrebno ispoštovati kako bi metoda dala zadovoljavajuće rezultate. Ti koraci su sljedeći:

1. Određivanje ciljeva analize grupa
2. Određivanje istraživačkih obrasca
3. Određivanje pretpostavki
4. Formiranje i procjena broja grupa
5. Interpretacija grupa
6. Procjena analize grupa i profiliranje grupa (Hair et al., 2010.)

Osim spomenute primjene analize grupa u medicini, analiza grupa pronalazi široku primjenu u gotovo svim znanostima kao i ostalim bitnim segmentima čovjekova djelovanja. Tako primjerice u sportskom svijetu točnije u nogometu jedna od poznatijih priča jest priča o vlasniku jednog omanjeg engleskog kluba pod nazivom Brentford. Naime vlasnik navedenog kluba odlučio je raznim statističkim analizama pokušati poboljšati svoj klub. Te je tako primjerice grupiranjem igrača za koje smatra da su mu potrebni ovisno o njihovim sposobnostima uspio izvući najbolje što je mogao. Ne samo da mu je takav pristup donio uspjeh već je doveo svoj klub do jednog od najvećih postignuća u povijesti engleskog sporta. Drugi primjer kako analiza grupa može dovesti do raznih postignuća jest kod procesa segmentacije tržišta. U tom procesu kreiraju se grupe potrošača kako bi se svakoj grupi ponudilo točno ono što bi ta grupa potrošača bila voljna kupiti, ali ne samo da im se ponudi ono što bi pojedinci htjeli kupovati već je i bitno da se analizira i njihova platežna moć kako ne bi došlo do ponude preskupih proizvoda ljudima koji jednostavno nisu u mogućnosti si priuštiti iste. Široku primjenu ova metoda pronalazi i u marketingu kako bi se pojedine grupe ljudi zainteresiralo za određene proizvode i usluge (Hair et al., 2010.).

U samoj analizi grupa postoji nekoliko metoda kojima se određuju sličnost među objektima. Točnije postoje tri metode, a one su kako slijedi: mjere udaljenosti, mjere korelacije, mjere udruživanja. Počevši od najraširenije mjere u praksi, točnije mjere udaljenosti. Za mjeru udaljenosti kažemo da je numerička mjera odnosno u toj mjeri koriste se numeričke vrijednosti varijabli. Postoji različite mjere udaljenosti pa tako imamo: Euklidova udaljenost, kvadrirana Euklidova udaljenost, Manhattan udaljenost, Mahalanobisova udaljenost. Te mjere nam pokazuju zapravo koliko su objekti udaljeni, odnosno različiti. Samim time što dobijemo veću numeričku vrijednost za mjeru udaljenosti to su objekti različitiji. Nadalje imamo mjere korelacije koje su također baš kao i mjere udaljenosti numeričkog tipa. Takve mjere temelje se

na određivanju koeficijenta korelacije između parova objekata izmjerenih na više varijabli. Što je koeficijent korelacije veći to je veća sličnost među objektima (Hair et al., 2010.).

Grupiranje podataka je jedna od iznimno važnih metoda kod gotovo svih ljudskih djelatnosti svoju širu primjenu može zahvaliti ubrzanom razvoju informatičkih tehnologija. Tako je analiza grupa svoj procvat doživjela tek šezdesetih godina prošlog stoljeća. Budući da svakodnevno svijetom razmjenjuje se velika količina podataka ljudski um nije u stanju procesirati toliko količinu podataka u relativno kratkom vremenu. Tu nastupaju računala koja sve većim razvojem ne samo da sve brže obrađuju sve veće količine informacija već ih i sve točnije obrađuju. Pa tako i analiza grupa paralelno razvoju informatičke tehnologije bilježi i svoj sve veći razvoj i sve veću primjenu (Nachtway et al., 2009.).

Od iznimne je važnosti ukazati na to kako su analiza grupa i klasifikacija dva različita pojma. Budući da ih mnogo ljudi smatram pod jednake potrebno je dobro objasniti razliku između istih. Klasifikacija za razliku od analize grupa unaprijed ima određen broj grupa te svojstva koja čine tu grupu. Dok se kod analize grupa ne zna niti konačni broj grupa niti koja svojstva su sadržana u pojedinoj grupi.

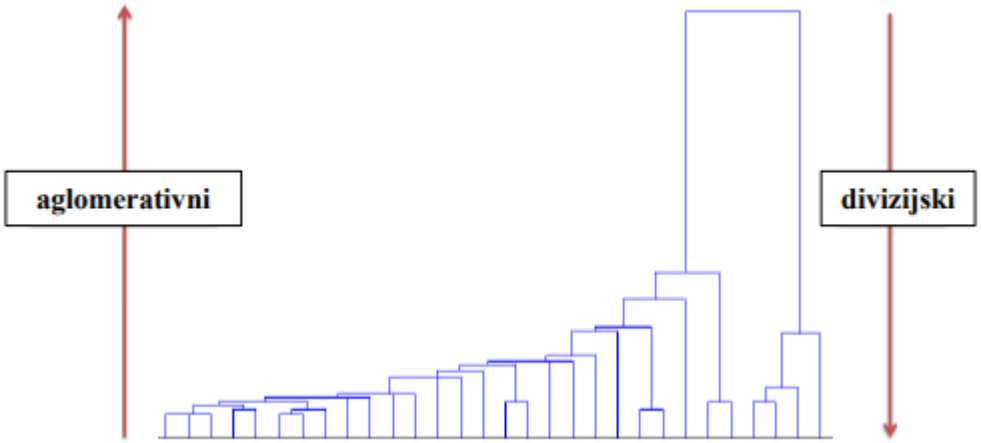
3.1. Podjela metoda za analizu grupa

Kako je već napomenuto postoji vrlo mnogo algoritama za provedbu analize grupa. No, u cijeloj toj gomili algoritama dva pristupa se izdvajaju po svojoj učinkovitosti. Jedan od tih pristupa naziva se hijerarhijski pristup. Hijerarhijski pristup, odnosno metoda kao svoj krajnji produkt ima dendrogram. Dendrogram jest grafički prikaz grupa u obliku stabla povezivanja. Hijerarhijska metoda započinje tako se izračuna udaljenost između svih objekata zajedno. Nakon što se izračuna njihova međusobna udaljenost kreće se u stvaranje grupa. Grupe se stvaraju jednom od dvije tehnika. Ovisno kako je moguće formirati grupe koristimo se spajanjem ili razdvajanjem. Tehnika spajanja ili aglomerativna metoda da se svaki objekt nalazi sam u zasebnoj grupi. Nakon što su svi objekti raspoređeni u svoje grupe proces kreće dalje. Uspoređivanjem grupa spajamo slične grupe koje formiraju sve veću grupu. Proces se nastavlja tako dugo dok se svi objekti na kraju ne nalaze u jednoj velikoj grupi. Drugi oblik naziva se divizijska metoda ili tehnika razdvajanja. Kod navedene metoda radi se potpuno suprotno od aglomerativne metoda. Dakle, polazišna točka je ta da su svi objekti svrstani u jednu veliku grupu. Iz te velike polazišne grupe gledamo koji su objekti dovoljno različiti da bi se svrstali u novu grupu. Tom metodom dobivamo iz jedne grupe dvije, pa iz nastale grupe još dvije i tako dalje sve dok se naposljetku ne dobije svaki objekt promatranja zasebno. Iako su obje metode zadovoljavajuće aglomerativna metoda koristi se mnogo češće od divizijske. Drugi pristup analize grupa jest nehijerarhijski pristup. Kod nehijerarhijskog pristupa potrebno

je unaprijed odabrati broj grupa. Određivanjem grupa dobiva se mogućnost prelaska objekata iz jedne grupe u drugu. Isto kao i hijerarhijski pristup i ovaj pristup pronalazi široku primjenu pa tako postoji i mnogo različitih varijanti istog. Iako postoji više različitih varijanti ovog pristupa u samoj svojoj srži svi oni koriste istu tehniku. Sama bit tehnike jest pronaći grupu oko koje se grupiraju objekti te izračunima pronaći nove točke grupiranja ovisno o prosječnoj vrijednosti objekata. Tokom tog procesa objekti putuju iz jedne grupu u drugu i tako sve dok se ne zadovolji uvjet stabilnosti za već određen broj grupa (Ekonomski fakultet Subotica, predavanje, 2015.).

3.1.1. Hijerarhijski pristup

Kako je već napomenuto hijerarhijski pristup može se podijeliti na dvije metode. Te dvije metode jesu aglomerativna metoda i divizijska metoda. Aglomerativna metoda se razvija tako da na početku su svi objekti razvrstan zasebno pa se spajaju u sve veće grupe ovisno o njihovoj sličnosti. Za razliku od aglomerativne metode, divizijska metoda polazi od stajališta da su svi objekti svrstani u jednu zajedničku grupu te se postepeno razvrstavaju u sve veći broj grupa dok na posljetku ne bude svaki objekt zasebno (Ekonomski fakultet Subotica, predavanje, 2015.).



Slika 1. Hijerarhijski pristup (Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Aglomerativna metoda započinje tako da se kreira matrica udaljenosti objekata. Na početku svi objekti imaju istu veličinu i ona iznosi jedan. Budući da, svaki objekt čini svoju zasebnu grupu gleda se koje su grupu blizu te ih se spaja. Spajanjem takvih grupa nastaju veće grupe koje sadrže više objekata odnosno podgrupa. Budući da je moguće na različite načine odrediti koje grupe su blizu potrebno se odlučiti za jednu od metoda. Najjednostavnija metoda za odabir blizine objekta je metoda kojom preko susjeda gledamo što je blizu. Kako se određuje što jest, a što nije prikazano je na primjeru ispod.

Tablica 1. Udaljenost grupa

	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-

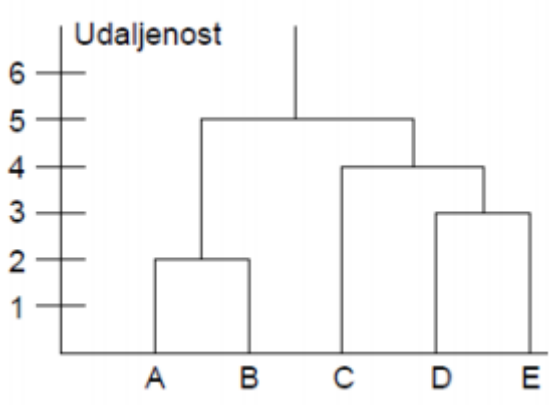
(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Tablica 2. Grupiranje grupa ovisno o blizini susjeda

Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
4	(A, B), (C, D, E)
5	(A, B, C, D, E)

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Gledajući udaljenost grupa spajamo one grupe kojima su objekti iz jedne grupe najbliži nekim objektima iz druge grupe. Kada nađemo koji objekti su najbliži tada njihove grupe možemo spojiti i formirati novu grupu. Na prikazanom primjeru možemo vidjeti kako se gledaju udaljenosti. Vidljivo jest da je najmanja udaljenost između grupa 2 te tako spajamo grupu A i B. Budući da, smo spojili grupu A i B sada imamo četiri grupe a to su redom: AB, C, D i E. U sljedećem koraku opet gledamo udaljenosti i vidimo da je najmanja udaljenost 3 između grupa D i E. Zbog toga spajamo grupe i dobivamo tri grupe: AB, C i DE. Proces opet ponavljamo te uviđamo da je najmanja udaljenost sada između grupa C i DE te spojimo i te grupe te sada dobivamo sljedeće grupe: AB i CDE. Krećemo u slijedeći korak i vidimo da je najmanja udaljenost između grupa AB i CDE pa spojimo i te grupe. Sada dobivamo grupu: ABCDE. U tom trenutku naš proces završava jer smo dobili jednu grupu koja sadrži sve ostale podgrupe. Na slici ispod prikazan je dendrogram koji prikazuje kako je došlo do stvaranja grupa.



Slika 2. Dendrogram grupiranja najbližih susjeda (Izvor: Ekonomski fakultet Subotica, predavanja 2015.)

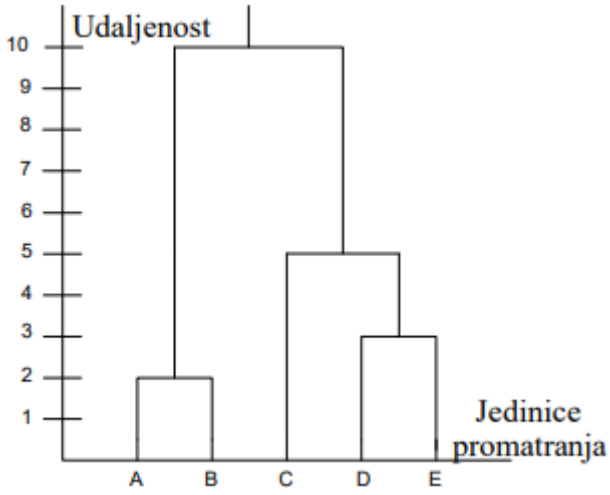
Druga vrlo raširena metoda jest povezivanje najdaljih susjeda. Koristit ćemo isti primjer za prikaz ovakve vrste spajanja.

Tablica 3. Grupiranje grupa prema najdaljim susjedima

Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
5	(A, B), (C, D, E)
10	(A, B, C, D, E)

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

U ovakvoj vrsti povezivanja, grupe se povezuju ako je udaljenost najdaljih objekata iz grupa najmanja udaljenost. Koristeći se tom logikom nakon što su se formirale grupe AB, C i DE, povezivanje grupa C i DE će se izvršiti na udaljenosti koja iznosi 5. Zato što je ta udaljenost jednako udaljenosti najudaljenijih objekata iz te dvije grupe. Dok će se u slijedećem koraku kod grupiranja grupa AB i CDE kreirati grupa od ABCDE, ali ovaj puta na udaljenosti od 10. Zbog toga što je koristeći se istom logikom udaljenost između najudaljenijih objekata tih grupa, odnosno objekta A i objekta D, jednaka 10. Dendrogram kreiranja grupa prikazan je na slici ispod.



Slika 3. Dendrogram grupiranja najudaljenijih susjeda (Izvor: Ekonomski fakultet Subotica, predavanja 2015.)

Još jedna tehnika koja pronalazi svoju široku primjenu jest povezivanje putem prosječnih udaljenosti među susjedima. Tehnika se bazira na tome da se izračunaju prosječne udaljenosti među svim grupama i tada se bira koja od prosječnih udaljenosti jest najmanja. Kada se odabrala najmanja prosječna udaljenost dvije grupe se spajaju u jednu novu grupu. Primjer ovakvog spajanja prikazan je na dolje prikazanom primjeru.

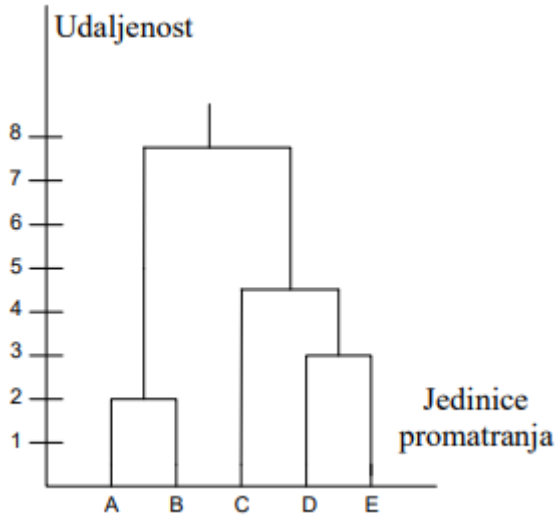
Tablica 4. Grupiranje grupa prema prosječnim udaljenostima

Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
4,5	(A, B), (C, D, E)
7,8	(A, B, C, D, E)

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

U ovom primjeru vidimo ponovno kako su prve dvije iteracije jednake no kod treće je sada vidljivo kako su grupe C i DE spojene na udaljenosti od 4,5. Isto tako je vidljivo kako su i u slijedećem koraku grupe AB i CDE spojene na udaljenosti od 7,8. Do toga je došlo tako što je izračunata prosječna udaljenost između grupa. I to tako što su uzete udaljenosti od objekta A do objekata C, D i E. Te udaljenosti iznose 6, 10 i 9. Isto tako uzeta je i udaljenost objekta B do objekata C, D i E. Te udaljenosti iznose 5, 9 i 8. Nakon što su uzete sve udaljenosti izračunat

je njihov prosjek na slijedeći način (6 + 10 + 9 + 5 + 8) / 6 što iznosi 7,8. Dendrogram ovog tipa spajanja prikazan je na slici ispod.



Slika 4. Dendrogram grupiranja tehnikom prosječne udaljenosti (Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

3.1.1.1. Metode povezivanja

Metode povezivanja dijele se na tri metode ovisno o tome kako se određuje predstavnik grupe. Te tri metode se sljedeće:

1. Jednostruko povezivanje *(engl. single-linkage method* ili *nearest-neighbor method)*
2. Potpuno povezivanje *(engl. complete-linkage metoda* ili *farhest-neighbor method)*
3. Prosječno povezivanje *(engl. average linkage)*

Kod metode jednostrukog povezivanja radi se o povezivanju najbližih susjeda ili minimalne udaljenosti. U toj metodi definira se sličnost između grupa kao najmanja udaljenost između bilo koja dva objekta iz tih dviju grupa. Grupe se stvaraju tako da se objekti čija je udaljenost najmanja povežu i formiraju grupu. Nakon kreiranja prve grupe sljedeći objekti se povezuju s tom grupom tako što se gleda najmanja udaljenost od samostalnih objekata i objekata već kreirane grupe. U narednim koracima postupak se ponavlja tako što se udaljenost grupa gleda na način da se uzima najmanja udaljenost između bilo koja dva objekta tih grupa (Kiš, 2012.).

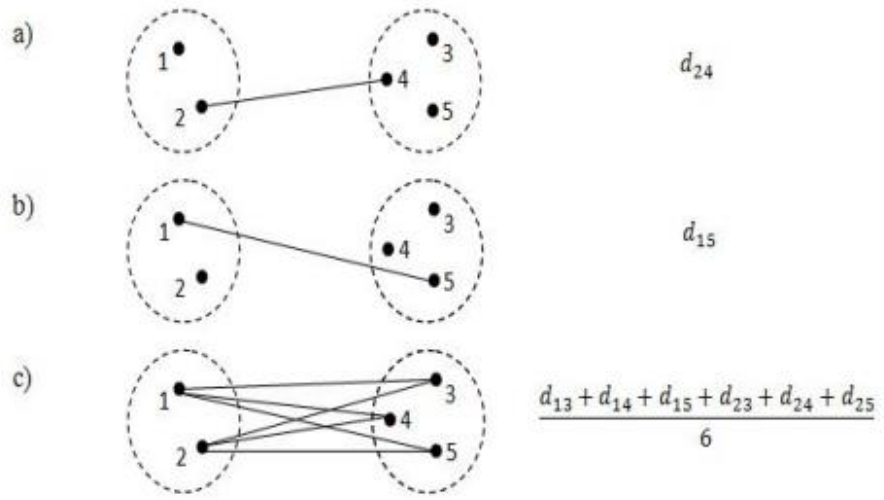
Iako je ovo vrlo široko primijenjena metoda postoji i veliki nedostaci iste. Kako ova metoda spaja grupe na način da gleda najkraće udaljenosti između grupa, nemoguće je odrediti koje su grupe loše razdvojene. Baš zbog toga što metoda ne može vidjeti loše razdvojene grupe može doći do nastanka ulančanih grupa. U ulančanim grupama objekti na krajevima grupa mogu biti potpuno različiti što na koncu dovodi do pogrešnih zaključaka. Zbog

mogućih takvih ishoda uvijek je potrebno provjeriti da li se uistinu radi o takvom rezultatu ili je jednostavno ova metoda neprihvatljiva za tu vrstu problema (Kiš, 2012).

Kod metode potpunog povezivanja radi se o povezivanju najudaljenijih susjeda ili maksimalne udaljenosti. Iako je ova metoda skoro pa jednaka metodi jednostrukog povezivanja, te dvije metode razlikuju se u jednoj ključnoj stvari. Dok metoda jednostrukog povezivanja povezuje grupe na način da traži najbližu vezu između pojedinih objekata u grupama, to kod metode potpunog povezivanja nije slučaj. Naime, u spomenutoj metodi grupe se povezuju tako što se gleda najudaljenija veza između pojedinih objekata. Ovim načinom osigurava se da je udaljenost između svih objekata grupa maksimalna. I samim time se izbjegava problem koji nastaje kod jednostrukog povezivanja (Kiš, 2012.).

Kod metode prosječnog povezivanja radi se o povezivanju grupa ovisno o tome kolika je prosječna udaljenost među objektima.

Na dolje slici prikazano je kako svaka od metoda funkcionira u praksi. Tako je vidljivo da jednostruko povezivanje spaja grupe temeljem njihovih najbližih objekata. S druge strane, potpuno povezivanje spaja grupe temeljem njihovih najudaljenijih članova. Dok se kod prosječnog povezivanja uzimaju u obzir svi elementi grupa te se računa prosječna udaljenost među objektima (Kiš, 2012.).



Slika 5. Udaljenost među grupama:

a) jednostruko povezivanje
b) potpuno povezivanje
c) prosječno povezivanje

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

3.1.1.2. Centroidna metoda

Ova metoda se zasniva na tome da se gleda udaljenost između centroida grupa kako bi se odradila sličnost među pojedinim grupama. Centroid klastera jest srednja vrijednost objekata u grupi po svim varijablama uključenima u analizu grupa. Svaki puta kada se dodaju novi objekti u grupu vrijednost centroida se mijenja. Navedena metoda vrlo je praktična i široko rasprostranjena u društvenim znanostima (Kiš, 2012.).

3.1.1.3. Warwadowa metoda

Ova metoda je najrazličitija od svih spomenutih metoda do sad. Osnovni princip ove metode jest analiza varijanca između objekata. Zbog toga što se u ovoj metodi analiziraju varijance metoda se još naziva i metoda minimalne varijance. Prilikom svakog spajanja grupa gleda se koje dvije grupe je porast ukupne sume kvadrata po svim varijablama u svim grupama minimalan. Kako bi se izračunala udaljenost u ovoj metodi najčešće se primjenjuje kvadrirana Euklidska udaljenost. Warwadowu metodu općenito je smatrana kao vrlo učinkovita. Cilj provedbe ovakve metode jest kreiranje grupa s malim brojem objekata i težnja da su grupe približno ujednačene što se tiče broja objekata (Kiš, 2012.).

3.1.2.Nehijerarhijske pristup

Kod nehijerarhijskog pristupa treba ranije definirati broj grupa . Ovisno o svom iskustvu, ranijih analiza ili preporuke sam znanstvenik koji provodi analizu određuje broj grupa koji smatra da će biti potreban. Nakon što odredi broj grupa koji želikoristiti, znanstvenik kreće u sljedeći korak koji se sastoji od razvrstavanja objekata. Postoji dva načina na koje se mogu objekti razvrstati u grupe. Jedan od načina jest da se slučajnim odabirom odaberu objekti koji će predstavljati točke grupiranja. Ostali objekti razvrstavaju se u grupe na način da se gleda udaljenost tih objekata od objekata koji predstavljaju točke grupiranja. Broj točaka grupiranja jednak je broju ranije definiranih grupa. Postupak premještanja objekata iz grupe u grupu ponavlja se nekoliko puta dok se ne dobiju grupe koje su što homogenije (Sebo et al., 2010.). Algoritam za provedbu nehijerahijskog pristupa izgleda ovako:

Određivanje točaka grupiranja

1. Pronalazak točaka unutar svake grupe na način da je udaljenost objekata što manja.
2. Pronađena točka definira se kao centroid. Lokacija cetroida uglavnom jest ne području najveće gustoće objekata.
3. Centroidi se postavljaju kao nove točke grupiranja budući da su optimalniji za točke grupiranja od točaka koje je znanstvenik sam postavio na početku algoritma.
4. Izračun udaljenosti svih objekata od centroida.
5. Određivanje novih centroida.

6. Ponavljanje postupka sve dok se ne formira homogeni sustav takav da je nemoguće ga poboljšati bilo kakvom drugom formacijom (Sabo et al., 2010.)

Metoda algoritma središnjih vrijednosti (engl. K-means metoda) dopušta da se objekti iz jedne grupu sele u drugu grupu. Baš taj proces preseljenja objekta iz jedne grupu u drugu razlikuje nehijerarhijski pristup od hijerarhijskog. Ime K-means proizlazi iz toga što se na samom početku određuje k elemenata koji predstavljaju točke grupiranja. U sljedećim koracima te točke zamjenjuju se centroidima (vektorima sredina) grupe (Sabo et al., 2010.).

3.1.3. Prednosti i nedostaci analize grupa

Jedna od najvećih prednosti analize grupa jest ta što se pojedine objekte svrstava u grupe te se svojstva objekata lakše odrede prema svojstvima cijele grupe. Sama analiza za rezultat ima minimalan gubitak informacija. Te se analizom grupa može vrlo jednostavno potvrditi ili opovrgnuti neke hipoteze koje biti možda bilo vrlo teško i/ili skupo dokazati na drugačiji način (Devčić et al., 2012.).

Postoje i neki nedostaci provedbe analize grupa. Jedan od nedostataka jest taj što nedostaje statističke osnove. Isto tako vrlo je lako moguće da znanstvenik subjektivno zaključuje o populaciji ovisno o testiranom uzorku. Zbog toga je važno odabrati vrlo pažljivo uzorak koji će se promatrati. Sam uzorak mora biti reprezentativan te što veći (Devčić et al., 2012.).

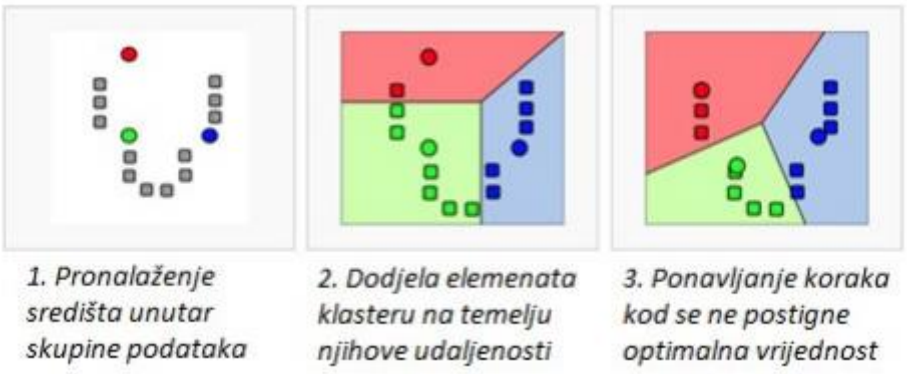
Još neki od nedostataka analize podataka su to što krajnje rješenje analize podataka neće uvijek biti jedinstveno već uvelike ovisi o tome koji pristup i metodu koristimo. Nadalje, analiza grupa uvijek kreira grupe neovisno o tome da li kreiranje grupa ima smisla ili ne. Isto tako znanstvenik mora voditi računa kada bira varijable po kojima će promatrati sličnost među objektima. Možda i najveći nedostatak analize grupa jest taj što uvelike ovisi o subjektivnim odlukama znanstvenika.

4. Algoritam k-središnjih vrijednosti *(k-means clustering)*

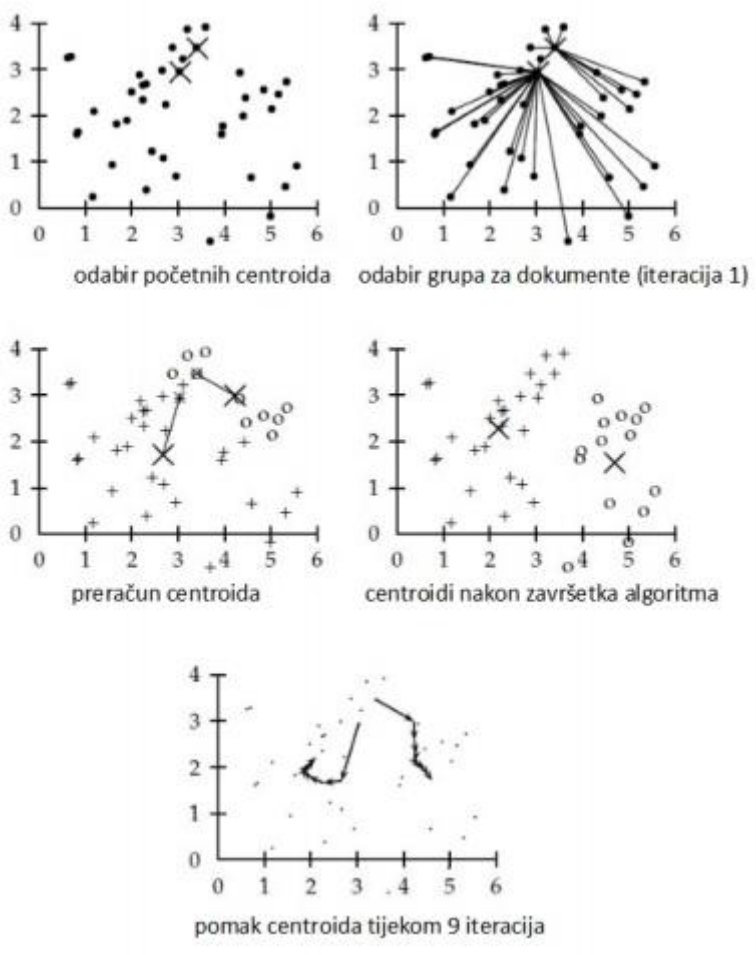
Osnovna ideja algoritma k-središnjih vrijednosti je određivanje predstavnika k skupina, i pridruživanje svake točke skupini s najbližim predstavnikom tako da zbroj kvadrata udaljenosti točaka od predstavnika skupina kojima pripadaju bude minimalan. Drugim riječima, algoritam k-središnjih vrijednosti kreira grupe koje su najkompaktnije, odnosno one grupe koje imaju najmanju varijaciju. Osnovni nedostatak ovog algoritam jest taj što kreira grupe koje su stabilne, ali ne uvijek i najoptimalnije. Još jedan nedostatak ovog algoritma jest taj šta su sve njegove skupine odvojive samo hiperravnima (Mirošević, 2016.).

Iako je jedan od problema algoritma optimizacija. Algoritam svoju primjenu pronalazi u pronalaženju particije koja je najbliža optimalnoj particiji. Algoritam jest jedan od najčešće korištenih algoritama baš kod rješavanja takvog problema budući da je to globalni problem te se primjena k-središnjih vrijednosti pokazala kao jedna od najboljih metoda pri rješavanju istog (Hair et al., 2010.).

Ovaj algoritam svoju srž pronalazi u koncentraciji. Broj grupa i ulazni parametar označuje se slovom *k*. Kada se odrede centri svaki objekt se dodjeljuje grupi. I to na način da se izračunaju udaljenosti objekata od centara grupa i svaki objekt se svrstava u grupu od čijeg centra je udaljenost tog objekta najmanja. Određivanje centara i grupiranje objekata u grupe se provodi tako dugo dok se ne dobiju stabilne grupe. Ovaj algoritam jest vrlo raširen budući da ga je vrlo jednostavno primijeniti u velikom broju problema te ga je isto tako vrlo jednostavno implementirati. Jedini uvjeti za njegovu implementaciju su da je moguće definirati udaljenost među objektima i da je *k* neka razumna vrijednost. Sam algoritam kao rezultat ima vrlo zadovoljavajuće rezultate.



Slika 6. Algoritam k-središnjih vrijednosti, k = 3 (Izvor: Hair et al., 2010.)

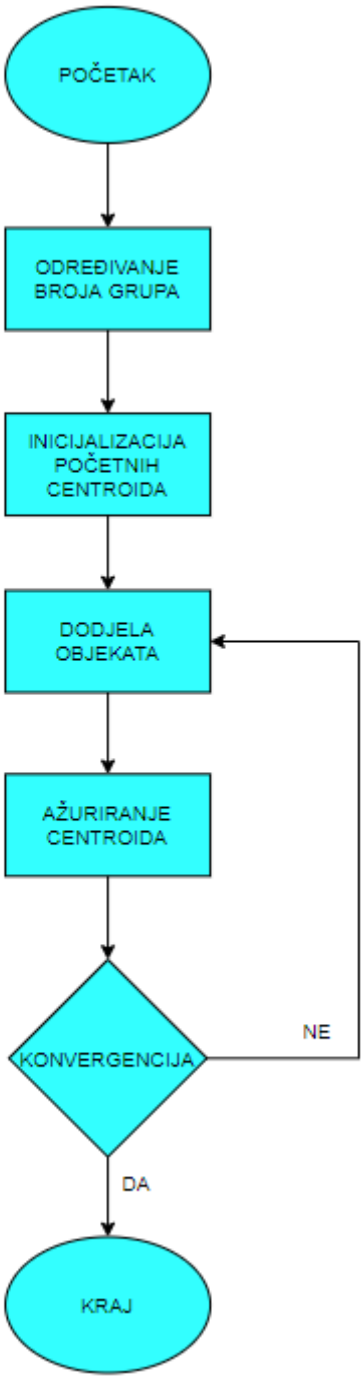


Slika 7. Prikaz rada algoritma k-središnjih vrijednosti (Izvor: Manning et al., 2009.)

Prema priloženoj slici iznad vidljivo je kako je algoritam k-središnjih vrijednosti vrlo jednostavan što se tiče samog poimanja istog. Definicija algoritma k-središnjih vrijednosti glasi: S obzirom na skup elemenata {x1, x2, ..., xn}, gdje je svaki element d-dimenzionalan vektor (d je prirodan broj), cilj je podijeliti skup u k grupa (Hair et. al., 2010).

Sam algoritam k-središnjih vrijednosti moguće je opisati kroz nekoliko koraka. Ti koraci su kako slijedi (Hair et al., 2010):

1. Inicijalizrianje *k* centroida.Kako bi bilo što jednostavnijeodabiru se slučajnim odabirom.
2. Pridruživanje objekta grupi s najbližim centoridom.
3. Izračun novih centroida grupa.
4. Ponavljanje koraka tako dugo dok se ne dobije stabilan sustav.



Slika 8. Dijagram algoritma k-središnjih vrijednosti (Izvor: izrada autora prema Hair et al., 2010.)

U sljedećem dijelu ovog poglavlja prikazat ću primjer kako se koristi algoritam k-središnjih vrijednosti. Za primjer ćemo uzet 8 točaka koje treba grupirati. Navedene točke su kako slijedi: (1, 1.5), (1.2, 1.3), (1.5, 1.8), (5, 5.01), (5.1, 4.7), (5.15, 5.3), (5.75, 5.2), (2, 6). Cilj jest grupirati ove točke u dvije grupe. Dolje navedena tablica pokazuje točke koje moramo grupirati.

Tablica 5. Primjer zadatka algoritma-k središnjih vrijednosti

Broj objekata	x	y
0	1	1.5
1	1.2	1.3
2	1.5	1.8
3	5.0	5.01
4	5.1	4.7
5	5.15	5.3
6	5.75	5.2
7	2.0	6.0

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Sada će biti prikazan postupak algortma k-središnjih vrijednosti. Prvi korak jest odrediti centroide grupa.

m_1 = (1.5, 1.8), m_2 = (5.0, 5.01)

Tablica 6 Početni centroidi

Grupa	Broj objekata	Centri
Grupa 1	2	m_1: (1.5, 1.8)
Grupa 2	3	m_2: (5.0, 5.01)

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Sljedeći korak zahtjeva izračunavanje euklidske udaljenosti između svij objekata i centroida. Dolje prikazana tablica prikazuje rezultate izračuna euklidske udaljenosti za svaki objekt posebno.

Tablica 7. Euklidska udaljenost rezultati kod prve iteracije

Broj objekata	Udaljenost do *m1*	Udaljenost do *m2*
0	**0.23570**	5.19255
1	**0.23570**	5.20570
2	**0.37712**	4.63221
3	5.12591	**0.46241**
4	4.99789	**0.73740**
5	5.43397	**0.55305**
6	5.81762	**1.15077**
7	4.53199	**2.70824**

(Izvor: Ekonomski fakultet Subotica, predavanja, 2015.)

Nakon što smo izračunali euklidske udaljenosti za svaki od objekata vrijeme je da se svaki objekt grupira u jednu od dviju grupa. Tim postupkom dobivamo sljedeće rezultate grupa1 = {0, 1, 2} i grupa 2 = {3, 4, 5, 6, 7}.

Sljedeći korak jest odrediti nove centroide. Kako bi odredili nove centroide računamo prosjek svih točaka unutar svake od grupa. Tako za grupu 1 imamo: m_1 = ((1 + 1.2 + 1.5), (1.5 + 1.3 + 1.8))/3 = (1.23333, 1.53333), dok za grupu 2 imamo: m_2 = ((5 + 5.1 + 5.15 + 5.75 + 2), (5.01 + 4.7 + 5.3 + 5.2 + 6))/5 = (4.6, 5.242). Nakon što smo dobili nove centroide ponavljamo postupak te ponovno izračunavamo euklidsku udaljenost za svaki objekt. Dobiveni rezultati prikazani su u tablici ispod.

Tablica 8. Euklidska udaljenost rezultati kod druge iteracije

Broj objekata	Udaljenost do *m1*	Udaljenost do *m2*
0	**0.23570**	5.19255
1	**0.23570**	5.20570
2	**0.37712**	4.63221
3	5.12591	**0.46241**
4	4.99789	**0.73740**
5	5.43397	**0.55305**
6	5.81762	**1.15077**
7	4.53199	**2.70824**

(Izvor: Ekonomski fakultet Subotica, 2015.)

Nakon što smo ponovno izračunali euklidsku udaljenost za svaki objekt ponovno grupiramo svaki objekt u grupe. Tim postupkom dobivamo slijedeće: grupa 1 = {0, 1 , 2} i grupa 2 = {3, 4, 5, 6, 7}.

Usporedimo li grupe iz prve i druge iteracije vidimo kako su dvije grupe jednake u obje iteracije. Time zaključujemo kako je algoritam k-središnjih vrijednosti došao do kraja.

Kao što smo mogli vidjeti algoritam k-središnjih vrijednosti vrlo dobro funkcionira u danom primjeru. No,to nije uvijek slučaj. Tako primjerice algoritam k-središnjih vrijednosti ima neka ograničenja. Ta ograničenja su sljedeća:

1. Ovisnost o početnim centroidima. Posebno primjetno kada grupe sadrže malo objekata.
2. Broj *k* mora se odrediti unaprijed i to jest jedna od najvećih mana danog algoritma. Određivanje optimalnog broja *k* jest najveći izazov kod korištenja algoritma k-središnjih vrijednosti.
3. Algoritmu k-središnjih vrijednosti nemogući je pronaći ne-sferne grupe (Ekonomski fakultet Subotica, predavanja, 2015.).

5. Analiza podataka u programu R

U današnje vrijeme kada se svi segmenti ljudske djelatnosti kompjuteriziraju, tako se i analiza podataka provodi preko računala. Budući da računalo može procesuirati mnogo više podataka u kraćem vremenu od ljudskog uma, sam proces prikupljanja podataka i njihova obrada uvelike je napredovala. Sukladno tome se i analiza grupa procesuira i provodi puno efikasnije i brže kako se razvija informacijska tehnologija (Institut Ruđer Bošković, priručnik 2020.).

Kako se razvijala tehnologija te analiza podataka u korak s razvojem tehnologija tako se i razvila nova računalna znanost. Ta nova računalna znanost koja se razvila primjenjuje računalne postupke i alate kako bi što efikasnije provela analizu podataka. Opisana znanost naziva se dubinska analiza podataka *(engl. dana mining)*. Budući da, je navedena znanost vrlo nova ne postoji jedinstveni pristup oko provedbe analize. Iako je ovo vrlo mlada znanost rapidno se razvija te je gotovo nemoguće popisati sve tehnike koje su se do sada razvile za analizu podataka (Institut Ruđer Bošković, priručnik 2020.).

Postoji neki određeni uvjeti koji bi trebali biti ispunjeni kako bi se analiza podataka pravilno provela, a pogotovo kada se govorio o korištenju računala. Nužan uvjet kako bi se provela analiza korištenjem računala jest da podaci budu prikazani tablično. Nije nužno, ali jest neko uobičajeno da se u tablici redovi prikazuju kao instance, dok se svaki stupac deklarira kao jedna od varijabli promatranog objekta. Kao rezultat analize podataka dobiva se tablica u kojoj su primjeri u redcima te je svaki od njih opisan atributima. Svi primjeri u tablici moraju biti opisani na potpuno jednak način te s istim redoslijedom atributa. Bitno je kod obrade podataka računalom da se rezultati analize prikazuju u obliku koje čovjek može razumjeti. Kod klasifikacijskog modeliranja razlikujemo dva postupka ovisno kako prikazuju podatke. Tako imamo postupke koji prikazuju podatke u obliku stabla odluke *(engl. decsion tree)* ili pravlia *(engl. rule)*. Ako pak se koristimo neklasificiranim postupcima tada se koristimo postupkom otkrivanja čestih uzoraka. Postoje i neki drugi postupci kako bi se prikazala analiza podataka te izvuklo neko znanje iz iste, ali to su uglavnom postupci koje čovjek vrlo teško, a u nekim slučajevima čak i nemoguće, može interpretirati. Jedan od zanimljivijih postupka jest postupak slučajnih šuma. Navedeni postupak vrlo je koristan te služi pronalaženju iznimka koje odstupaju od pravila. Iako cilj analize podataka nije otkrivanje izuzetaka, već analiza postojećih te grupacija istih u skupine kako bi se moglo odrediti neka zajednička obilježja, otkrivanje izuzetaka može vrlo znatno pomoći u spoznaji novih saznanja (Institut Ruđer Bošković, priručnik 2020.).

Budući da je danas ponuda računalnih programa veća nego ikada postoji i jako mnogo alata kojima se može odraditi analiza podataka. Za potrebe istraživanja najpogodnija su ona koja su

besplatna i u kojima se može puno toga prilagođavati. Jedan od takvih alatajest i R. R je program koji je vrlo jednostavno koristiti. Uz neko osnovno znanje iz programiranja i naravno dobrim poznavanjem statistike može se obraditi velika količina podataka u vrlo kratkom vremenu.

5.1. Primjer algoritma k-središnjih vrijednosti u programu R

Kao cilj ovog istraživanja postavljeno je otkrivanje veza između podataka. Istraživanje je provedeno nad podacima o kemijskoj strukturi vina. Podaci govore o kemijskoj strukturi vina koje je proizvedeno u istoj regiji, ali su različiti proizvođači. Odabrane varijable koje se koriste u istraživanju su slijedeće: alkohol, jabučna kiselina, alkalnost pepela, magnezij, fenoli te intenzitet boje. Ukupno se promatra 6 varijabli. Sami podaci sastoje se od 178 objekata što je respektabilan broj instanci za ovakvo istraživanje. Podaci su preuzeti s UCI Machine Learning Repository i dostupni su na slijedećem linku: https://archive.ics.uci.edu/ml/datasets/Wine.

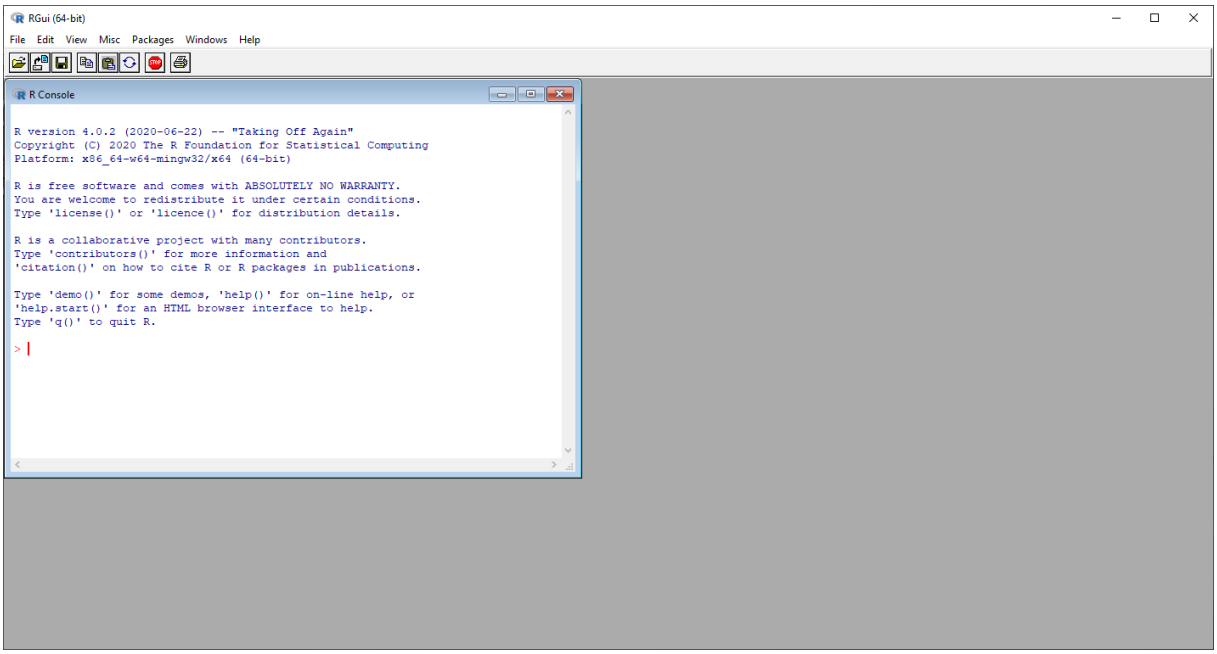
Tablica 9. Popis varijabli

Naziv varijable	Opis varijable	Vrsta varijable	Vrijednost podataka
Alcohol	Udio alkohola u vinu	Numerički	Float
Malic.acid	Jabučna kiselina	Numerički	Float
Acl	Alkalnost pepela	Numerički	Float
Mg	Magnezij	Numerički	Integer
Phenols	Fenoli	Numerički	Float
Color.int	Intenzitet boje	Numerički	Float

(Izvor: izrada autora)

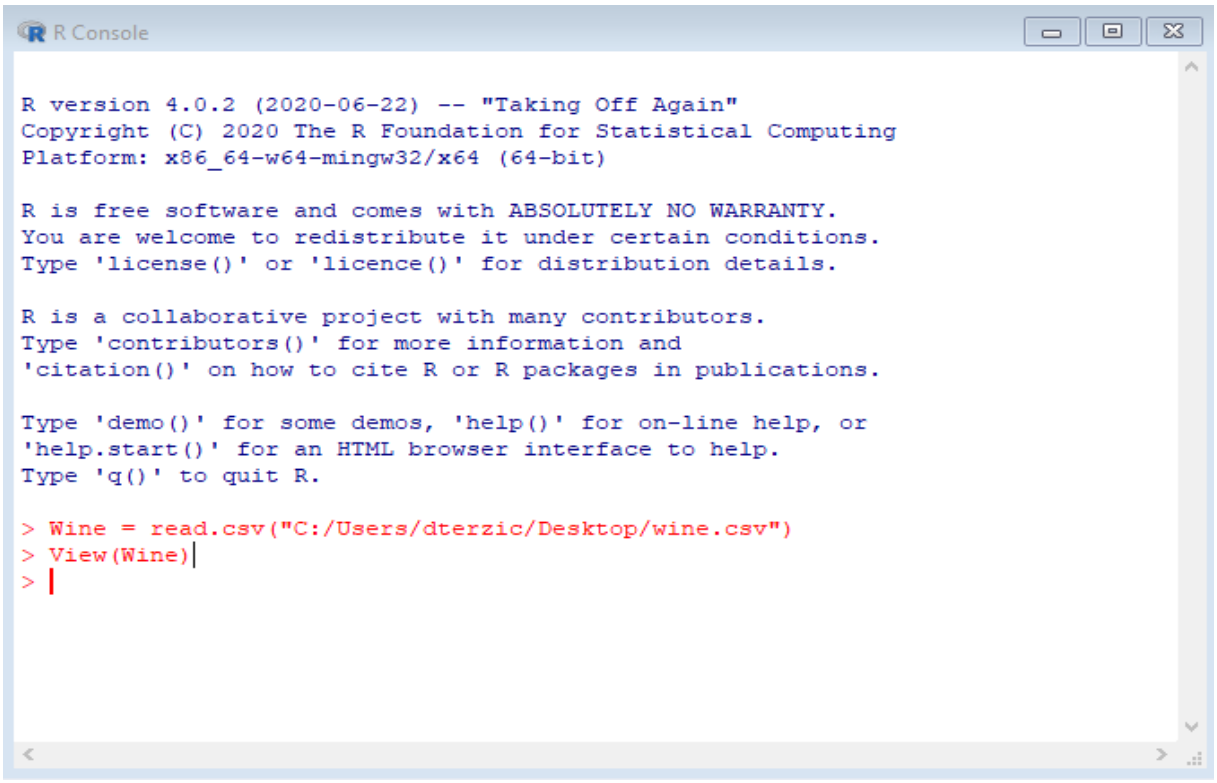
5.1.1. Obrada podataka

Prije same obrade podataka potrebno je provjeriti da li su svi podaci uneseni. Ukoliko postoje neki podaci kojima nedostaje vrijednost algoritam k-središnjih vrijednosti u R-u neće biti u mogućnosti obraditi podatke. Isto tako potrebno je da su podaci numerički. Prilikom pokretanja same aplikacije otvara se konzola kako je prikazano na slici ispod.

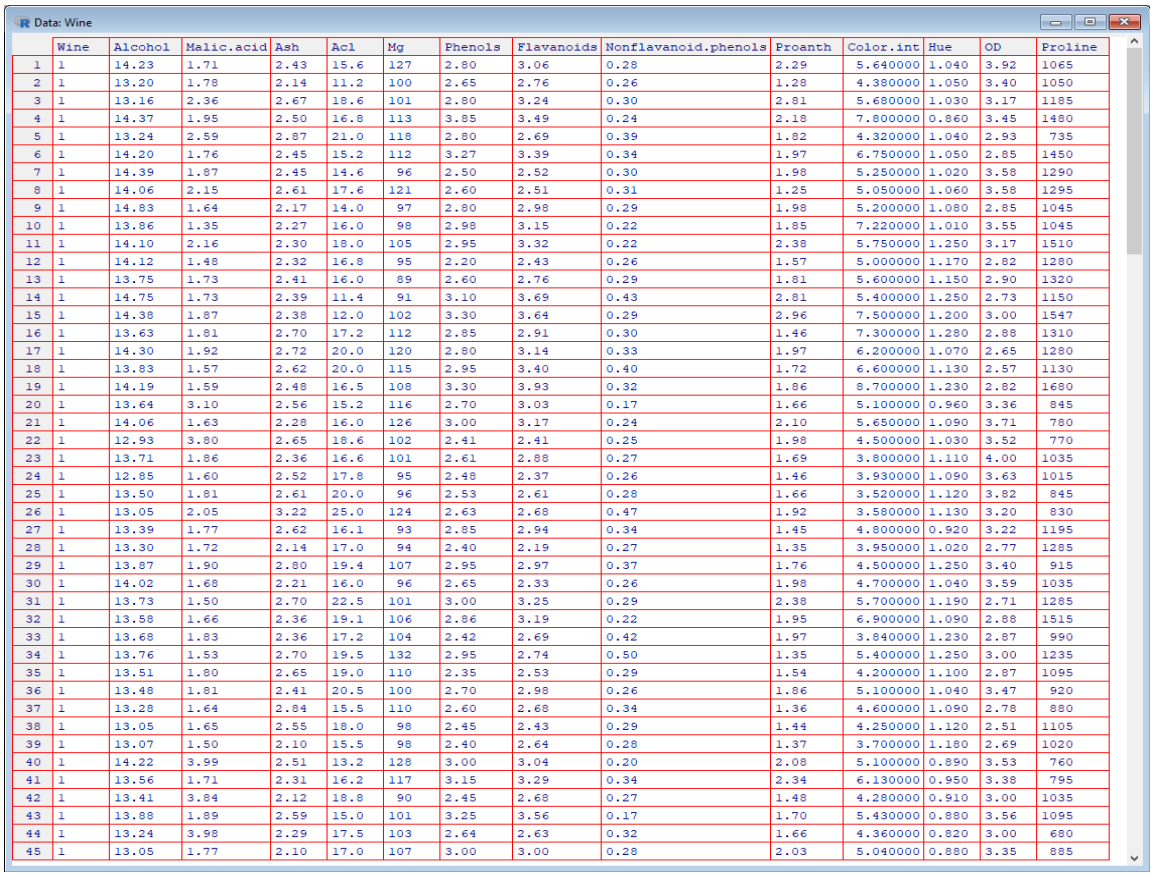


Slika 9. Početni ekran u R-u (Izvor: ekran autora)

Nadalje upisujemo u konzolu naredbe. Prva naredba koju je potrebno upisati jest naredba kako bi se učitala datoteka oblika .csv u kojoj se nalaze podaci za obradu. Nakon upisa naredbe za učitavanje datoteke može se upisati naredba koja će ispisati sve podatke iz baze podataka kako bi se uvjerilo da je sve prošlo u redu. Navedene operacije prikazane su na sljedećim slikama.



Slika 10. Učitavanje baze u R (Izvor: ekran autora)

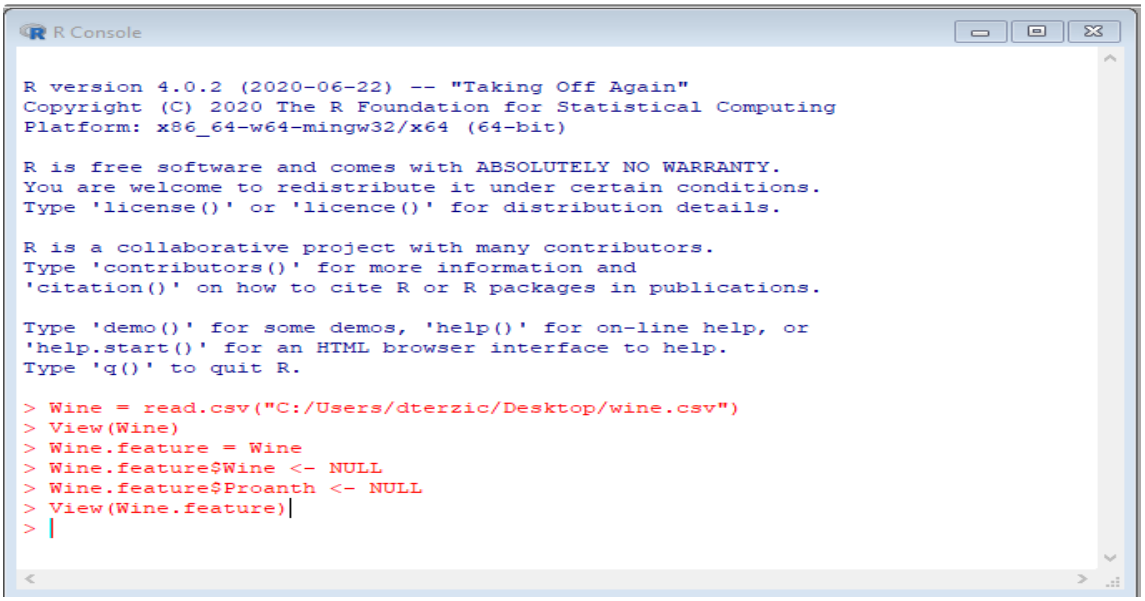


Data: Wine

	Wine	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue	OD	Proline
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040	3.92	1065
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050	3.40	1050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030	3.17	1185
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860	3.45	1480
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040	2.93	735
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.750000	1.050	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.250000	1.020	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.050000	1.060	3.58	1295
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.200000	1.080	2.85	1045
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.220000	1.010	3.55	1045
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.750000	1.250	3.17	1510
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.000000	1.170	2.82	1280
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.600000	1.150	2.90	1320
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.400000	1.250	2.73	1150
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.500000	1.200	3.00	1547
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.300000	1.280	2.88	1310
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.200000	1.070	2.65	1280
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.600000	1.130	2.57	1130
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.700000	1.230	2.82	1680
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.100000	0.960	3.36	845
21	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10	5.650000	1.090	3.71	780
22	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98	4.500000	1.030	3.52	770
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.800000	1.110	4.00	1035
24	1	12.85	1.60	2.52	17.8	95	2.48	2.37	0.26	1.46	3.930000	1.090	3.63	1015
25	1	13.50	1.81	2.61	20.0	96	2.53	2.61	0.28	1.66	3.520000	1.120	3.82	845
26	1	13.05	2.05	3.22	25.0	124	2.63	2.68	0.47	1.92	3.580000	1.130	3.20	830
27	1	13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45	4.800000	0.920	3.22	1195
28	1	13.30	1.72	2.14	17.0	94	2.40	2.19	0.27	1.35	3.950000	1.020	2.77	1285
29	1	13.87	1.90	2.80	19.4	107	2.95	2.97	0.37	1.76	4.500000	1.250	3.40	915
30	1	14.02	1.68	2.21	16.0	96	2.65	2.33	0.26	1.98	4.700000	1.040	3.59	1035
31	1	13.73	1.50	2.70	22.5	101	3.00	3.25	0.29	2.38	5.700000	1.190	2.71	1285
32	1	13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95	6.900000	1.090	2.88	1515
33	1	13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97	3.840000	1.230	2.87	990
34	1	13.76	1.53	2.70	19.5	132	2.95	2.74	0.50	1.35	5.400000	1.250	3.00	1235
35	1	13.51	1.80	2.65	19.0	110	2.35	2.53	0.29	1.54	4.200000	1.100	2.87	1095
36	1	13.48	1.81	2.41	20.5	100	2.70	2.98	0.26	1.86	5.100000	1.040	3.47	920
37	1	13.28	1.64	2.84	15.5	110	2.60	2.68	0.34	1.36	4.600000	1.090	2.78	880
38	1	13.05	1.65	2.55	18.0	98	2.45	2.43	0.29	1.44	4.250000	1.120	2.51	1105
39	1	13.07	1.50	2.10	15.5	98	2.40	2.64	0.28	1.37	3.700000	1.180	2.69	1020
40	1	14.22	3.99	2.51	13.2	128	3.00	3.04	0.20	2.08	5.100000	0.890	3.53	760
41	1	13.56	1.71	2.31	16.2	117	3.15	3.29	0.34	2.34	6.130000	0.950	3.38	795
42	1	13.41	3.84	2.12	18.8	90	2.45	2.68	0.27	1.48	4.280000	0.910	3.00	1035
43	1	13.88	1.89	2.59	15.0	101	3.25	3.56	0.17	1.70	5.430000	0.880	3.56	1095
44	1	13.24	3.98	2.29	17.5	103	2.64	2.63	0.32	1.66	4.360000	0.820	3.00	680
45	1	13.05	1.77	2.10	17.0	107	3.00	3.00	0.28	2.03	5.040000	0.880	3.35	885

Slika 11. Ispis baze (Izvor: ekran autora)

Zbog povećeg broja entiteta u bazi na slici iznad nisu prikazani svi podaci iz baze. Ukoliko je potrebno izbaciti neke od obilježja za koja se ne želi raditi analiza, vrlo jednostavno se to učini u R-u. Što je točno potrebno napraviti prikazano je na slici ispod.



Slika 12. Nepotrebne varijable (Izvor: ekran autora)

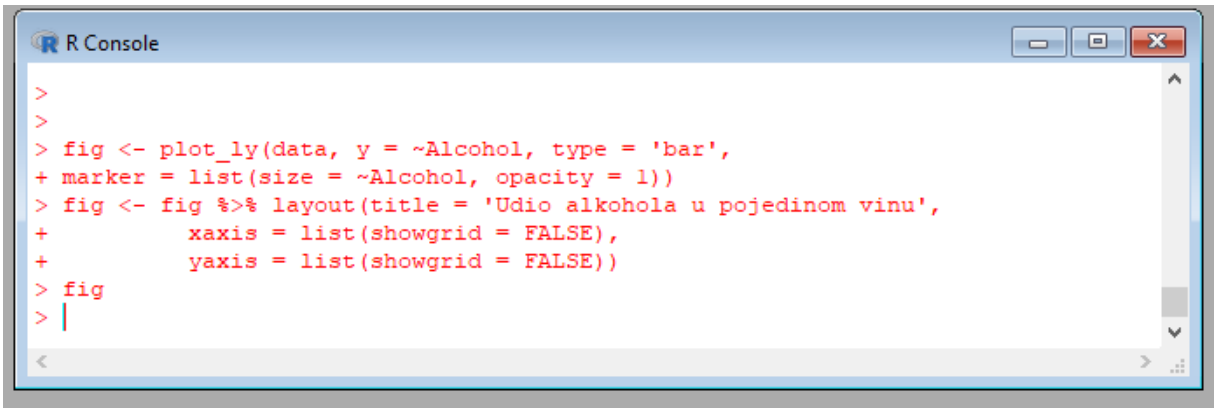
Budući da se ovim dvjema novim naredbama postiglo to da su iz baze maknute varijable koje ne želimo analizirati prikaz novih zapisa nalazi se na slici 13.

	Alcohol	Malic.acid	Acl	Mg	Phenols	Nonflavanoid.phenols	Color.int
1	14.23	1.71	15.6	127	2.80	0.28	5.640000
2	13.20	1.78	11.2	100	2.65	0.26	4.380000
3	13.16	2.36	18.6	101	2.80	0.30	5.680000
4	14.37	1.95	16.8	113	3.85	0.24	7.800000
5	13.24	2.59	21.0	118	2.80	0.39	4.320000
6	14.20	1.76	15.2	112	3.27	0.34	6.750000
7	14.39	1.87	14.6	96	2.50	0.30	5.250000
8	14.06	2.15	17.6	121	2.60	0.31	5.050000
9	14.83	1.64	14.0	97	2.80	0.29	5.200000
10	13.86	1.35	16.0	98	2.98	0.22	7.220000
11	14.10	2.16	18.0	105	2.95	0.22	5.750000
12	14.12	1.48	16.8	95	2.20	0.26	5.000000
13	13.75	1.73	16.0	89	2.60	0.29	5.600000
14	14.75	1.73	11.4	91	3.10	0.43	5.400000
15	14.38	1.87	12.0	102	3.30	0.29	7.500000
16	13.63	1.81	17.2	112	2.85	0.30	7.300000
17	14.30	1.92	20.0	120	2.80	0.33	6.200000
18	13.83	1.57	20.0	115	2.95	0.40	6.600000
19	14.19	1.59	16.5	108	3.30	0.32	8.700000
20	13.64	3.10	15.2	116	2.70	0.17	5.100000
21	14.06	1.63	16.0	126	3.00	0.24	5.650000
22	12.93	3.80	18.6	102	2.41	0.25	4.500000
23	13.71	1.86	16.6	101	2.61	0.27	3.800000
24	12.85	1.60	17.8	95	2.48	0.26	3.930000
25	13.50	1.81	20.0	96	2.53	0.28	3.520000
26	13.05	2.05	25.0	124	2.63	0.47	3.580000
27	13.39	1.77	16.1	93	2.85	0.34	4.800000
28	13.30	1.72	17.0	94	2.40	0.27	3.950000
29	13.87	1.90	19.4	107	2.95	0.37	4.500000
30	14.02	1.68	16.0	96	2.65	0.26	4.700000
31	13.73	1.50	22.5	101	3.00	0.29	5.700000
32	13.58	1.66	19.1	106	2.86	0.22	6.900000
33	13.68	1.83	17.2	104	2.42	0.42	3.840000
34	13.76	1.53	19.5	132	2.95	0.50	5.400000
35	13.51	1.80	19.0	110	2.35	0.29	4.200000
36	13.48	1.81	20.5	100	2.70	0.26	5.100000
37	13.28	1.64	15.5	110	2.60	0.34	4.600000
38	13.05	1.65	18.0	98	2.45	0.29	4.250000
39	13.07	1.50	15.5	98	2.40	0.28	3.700000
40	14.22	3.99	13.2	128	3.00	0.20	5.100000
41	13.56	1.71	16.2	117	3.15	0.34	6.130000
42	13.41	3.84	18.8	90	2.45	0.27	4.280000
43	13.88	1.89	15.0	101	3.25	0.17	5.430000
44	13.24	3.98	17.5	103	2.64	0.32	4.360000
45	13.05	1.77	17.0	107	3.00	0.28	5.040000
46	14.21	4.04	18.9	111	2.85	0.30	5.240000

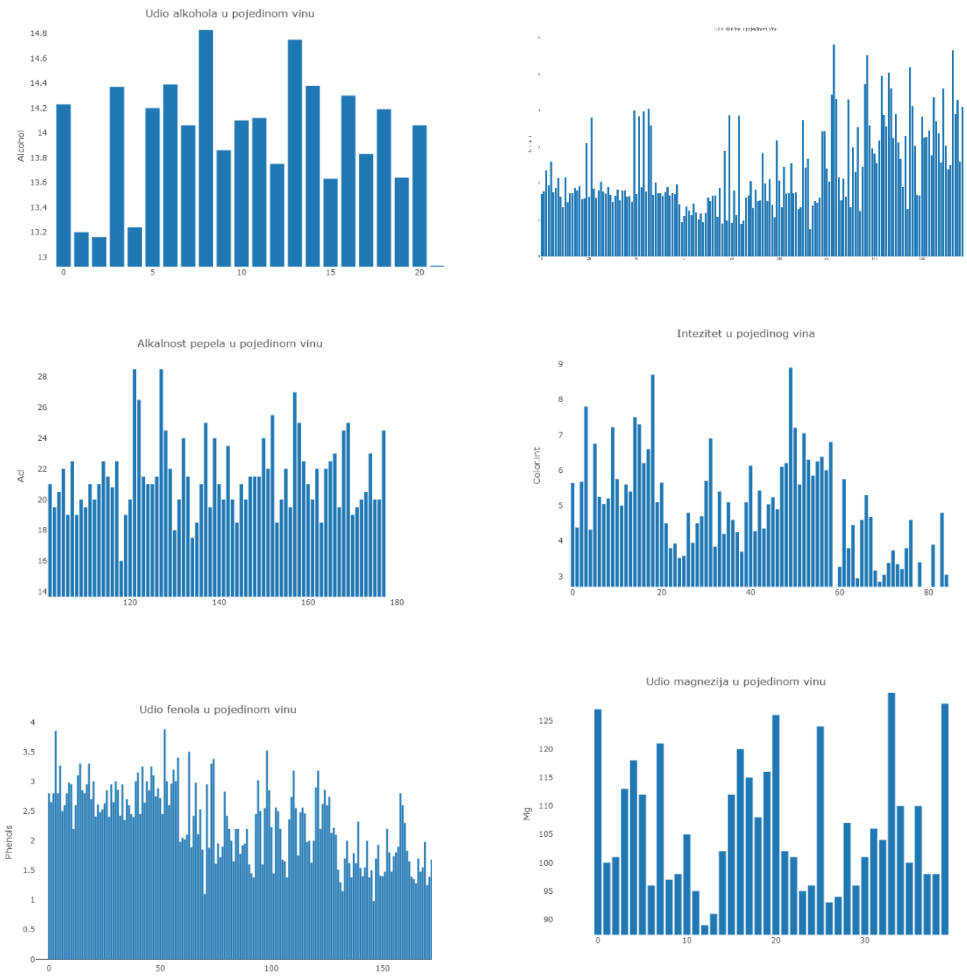
Slika 13. Baza bez varijabli (Izvor: ekran autora)

5.1.2. Rezultati

Nakon što se atributi odaberu moguće ih je grafički prikazati uz par jednostavnih linija koda. U nastavku je prikazan jedan primjer konzole kako se grafički mogu prikazati podaci te nekoliko grafova ovisno o varijabli koju želimo prikazati.

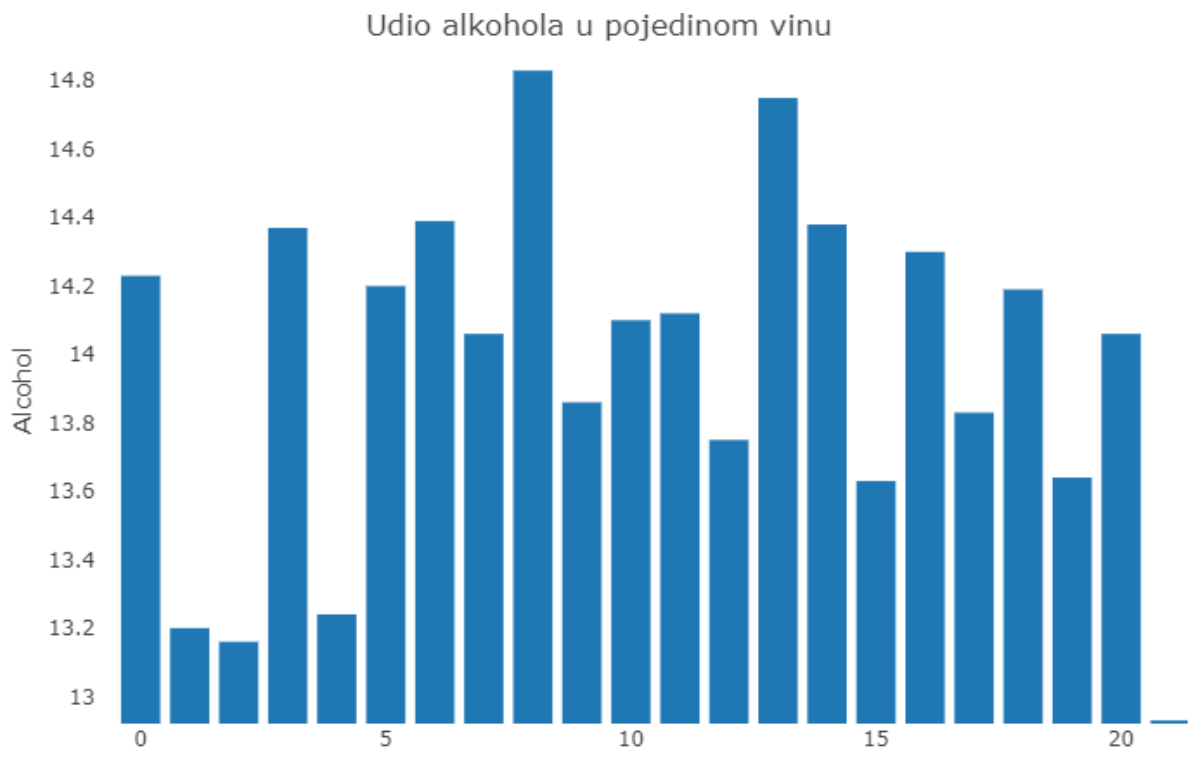


Slika 14. Prikaz konzole za ispis grafa (Izvor: ekran autora)



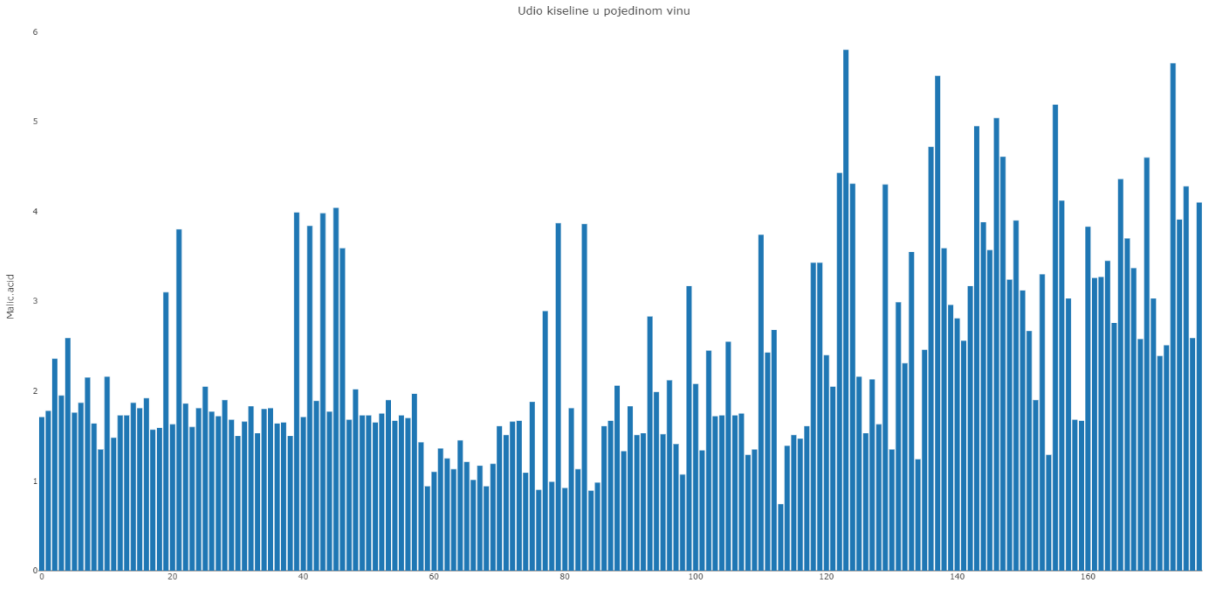
Slika 15. Prikaz grafova kreiranih u R-u (Izvor: ekran autora)

Slika 16. jest prikaz udjela alkohola u pojedinom vinu. Vidljivo je kako svaki objekt varira u odnosu na drugi.



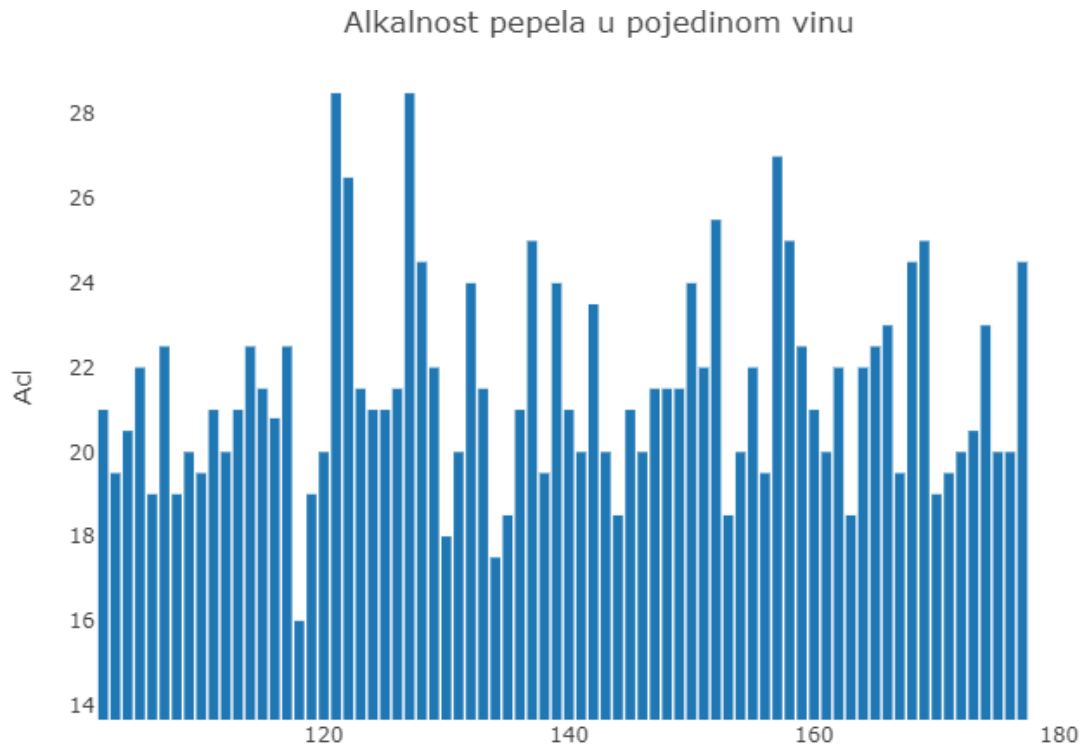
Slika 16. Udio alkohola u vinu (izvor: ekran autora)

Slika 17. jest prikaz udjela jabučne kiseline u pojedinom vinu.



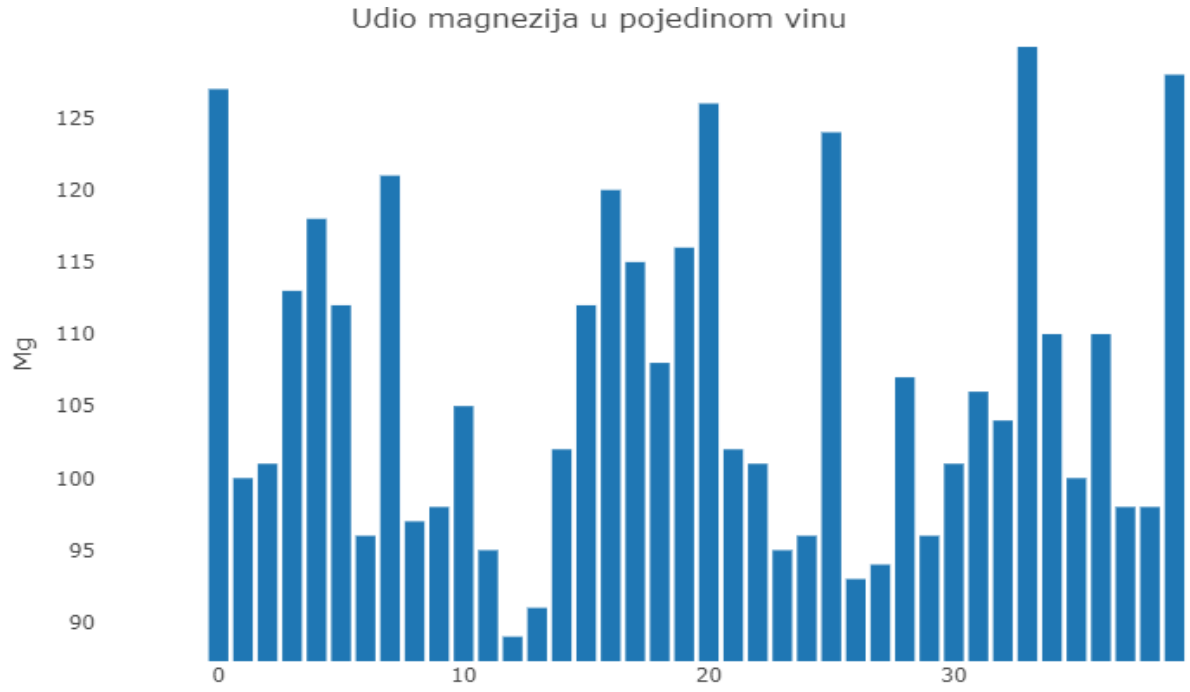
Slika 17. Udio kiseline (Izvor: ekran autora)

Slika 18. prikazuje alkalnost pepela u pojedinom vinu.



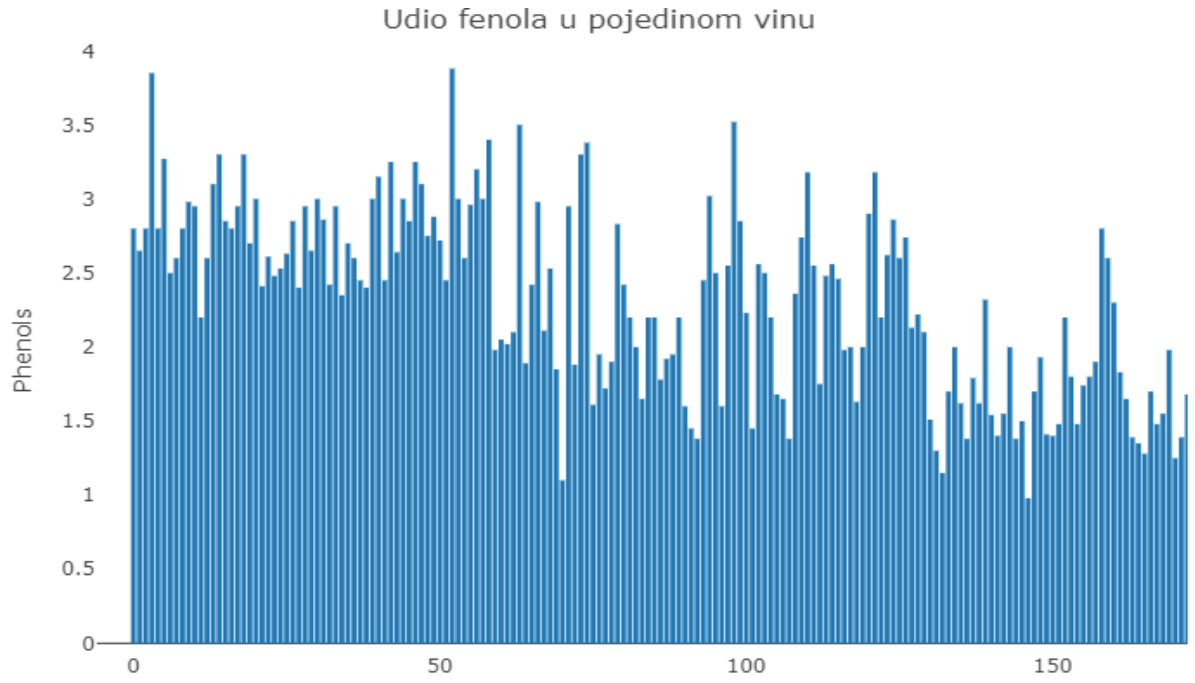
Slika 19. Alkalnost pepela (Izvor: ekran autora)

Slika 20. prikazuje udio magnezija u pojedinom vinu.



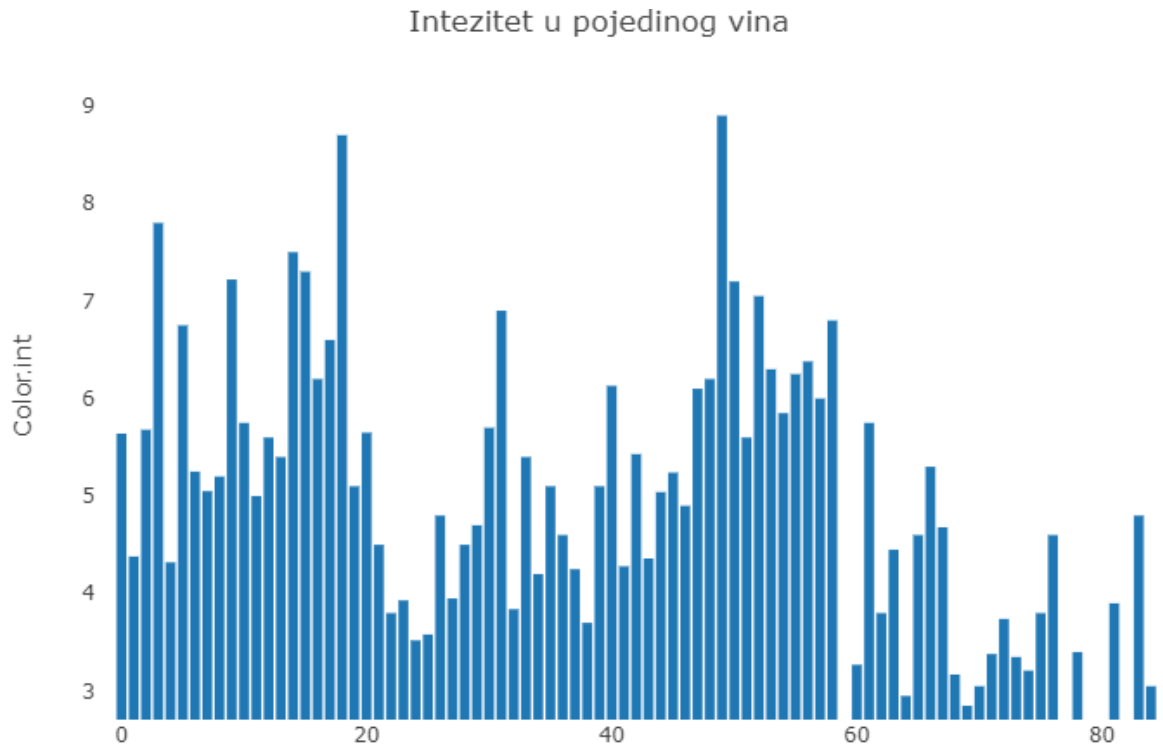
Slika 20. Udio magnezija (Izvor: ekran autora)

Slika 21. prikazuje udio fenola u pojedinom vinu.



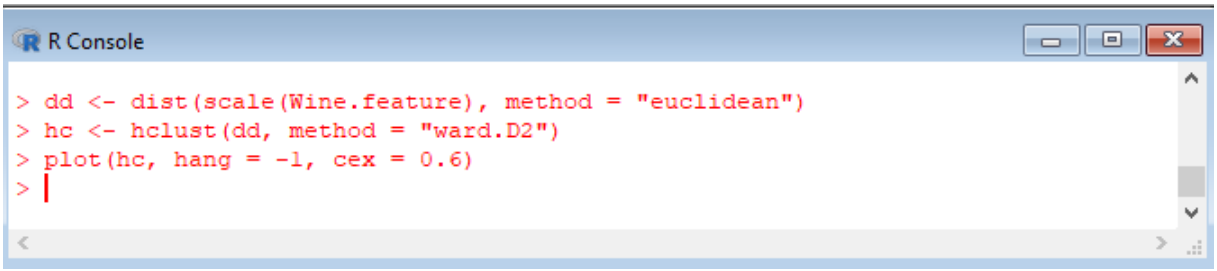
Slika 21. Udio fenola (Izvor: ekran autora)

Slika 22. prikazuje intenzitet boje pojedinog vina.

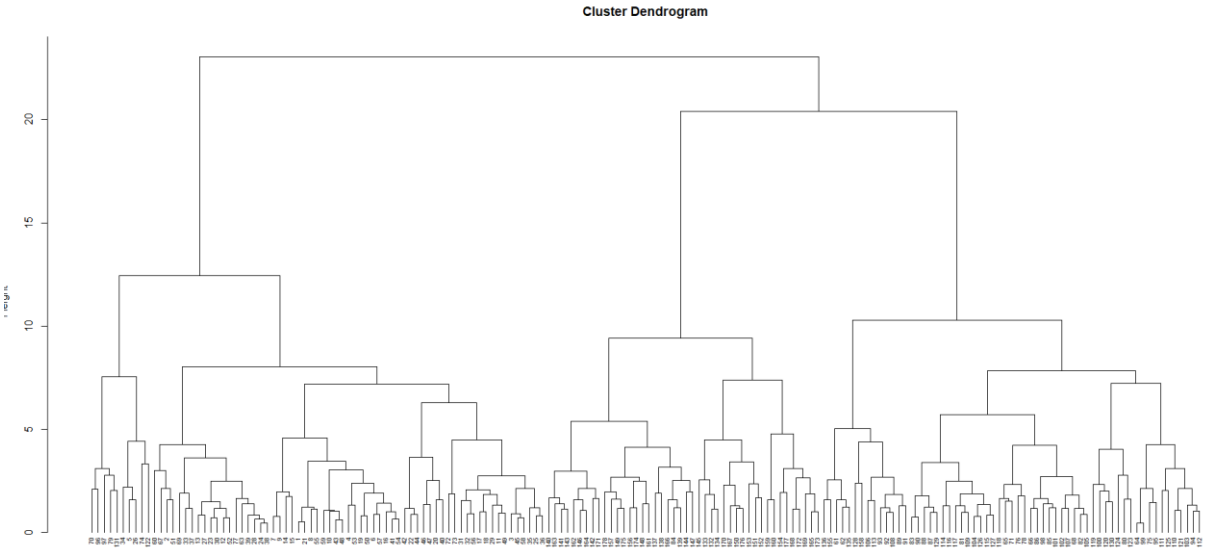


Slika 22. Intenzitet boje (Izvor: ekran autora)

Kao što je prikazano aplikacija R vrlo je korisna za mnogo toga kao primjerice izradu grafova. Osim za izradu grafova u ovom radu još je izrađen i dendrogram. Dendrogram je prikazan na slici 24. dok je kod za izradu istog prikazan na slici 23. Za izradu dendrograma korištena je euklidska udaljenost.

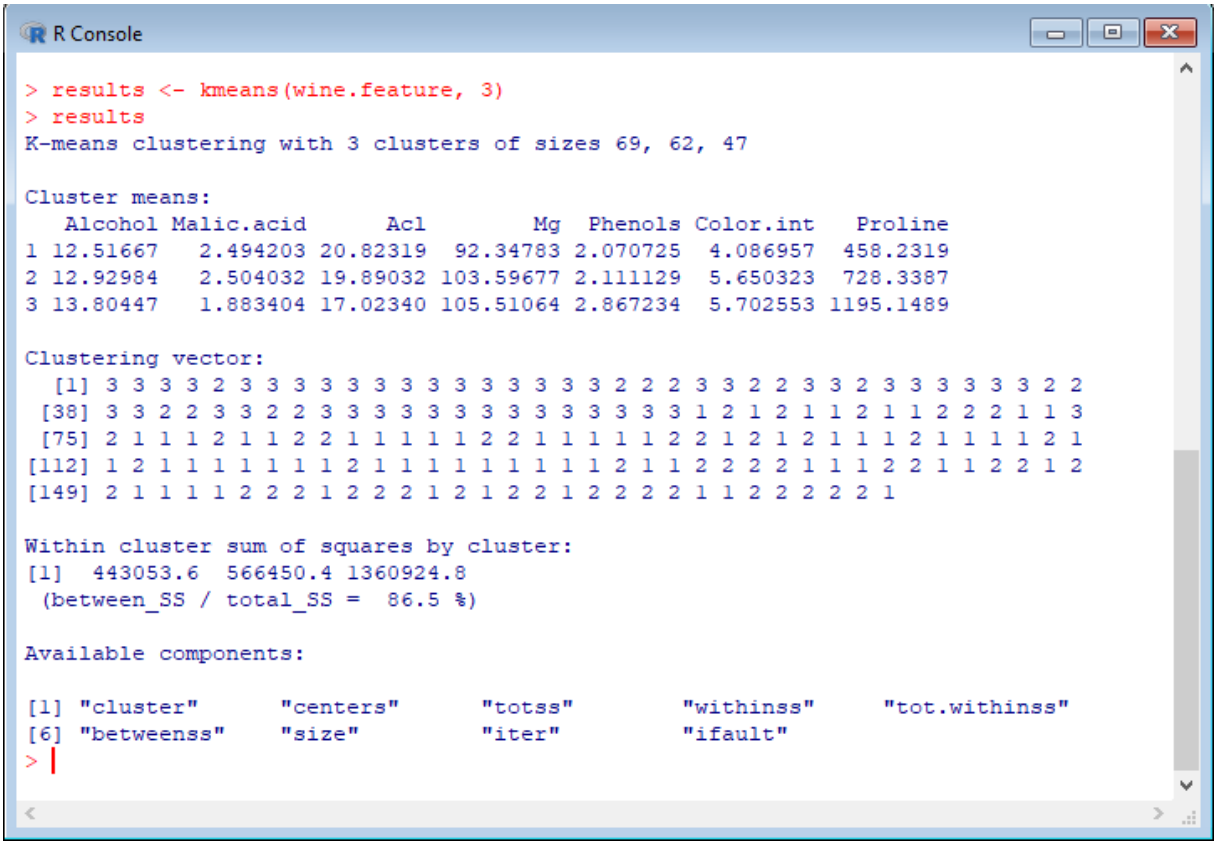


Slika 23. Kod za izradu dendrograma (Izvor: ekran autora)



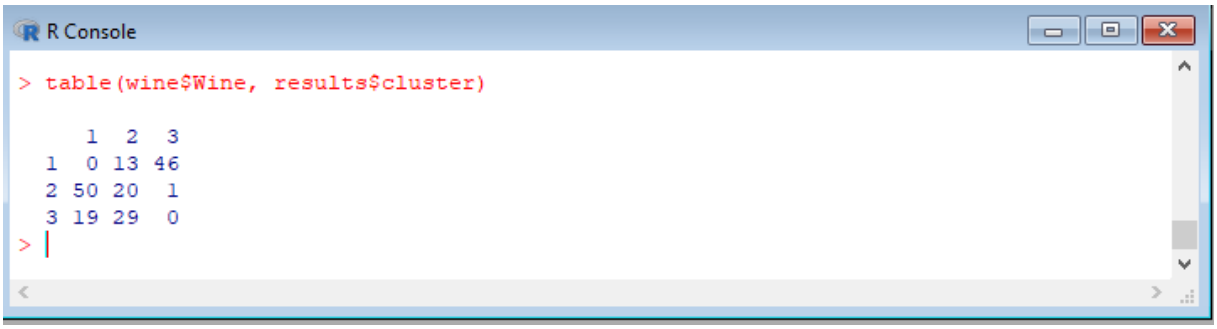
Slika 24. Dendrogram (Izvor: ekran autora)

Sljedeći korak jest provedba algoritma k-središnjih vrijednosti. Sam algoritam će grupirati podatke u 3 grupe. Za svaki od atributa proveden je algoritam k-središnjih vrijednosti. Rezultati su prikazani grafički. Objekti se grupiraju prema objektima u tri grupe kako bi se odredilo iz koje vinarije dolaze. Slika 25. prikazuje kod koji se koristi za provedbu algoritma k-središnjih vrijednosti kao i rezultate koji se dobiju provedbom istog.



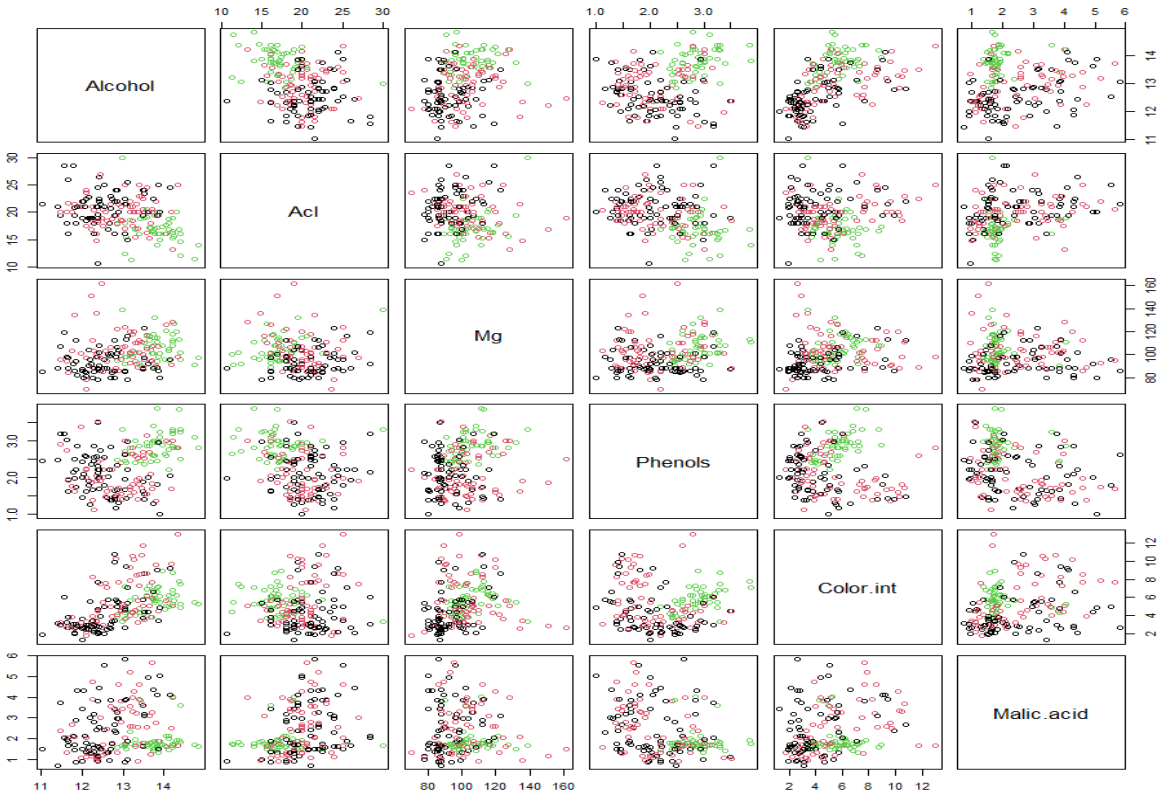
Slika 25. Algoritam k-središnjih vrijednosti (Izvor: ekran autora)

Iz rezultata može se iščitati kako su objekti raspoređeni u tri grupe te veličina pojedinih grupa. Iz slike konzole može se vidjeti kako prva grupa sadrži 69 objekata, druga 62, dok treća grupa sadrži 49 objekata. Isto tako vidljivo je i koliko iznose centroidi za pojedinu varijablu. Osim što se može vidjeti navedeni podaci vrlo korisno je što aplikacija R omogućuje i usporedbu između originalnih podataka i dobivenih rezultata. Tako je na slici 26. prikazan kod i ispis u konzoli koji prikazuje kolika su odstupanja u provođenju algoritma.



Slika 26. Usporedba rezultata (Izvor: ekran autora)

Iz slike je vidljivo kako su vina iz vinarija raspoređena u klastere. Algoritam k-središnjih vrijednosti rasporedio je vina iz prve vinarije tako što je stavio 13 vina u klaster broj 2 i 46 vina u klaster broj 3. Dok je vina iz vinarije 2 svrstao tako što je 50 vina stavio u klsater broj 1, 20 u klaster broj 2 te 1 vino u klaster broj 3 i na posljetku može se iščitati kako je vina iz vinarije svrstao tako što je stavio 10 u prvi klaster, 29 u drugi, dok u treći klaster nije stavio niti jedno vino. Prema ovim podacima vidljivo je kako je algoritam odradio zadovoljavajući posao. Iako nije u potpunosti točan, odstupanja su i očekivana budući da nisu korištene sve dostupne varijable već samo neke odabrane. Osim samog numeričkog i tabličnog prikaza rješenje R je pogodan i za grafičko prikazivanje istih. Tako je na slici 27. grafičk prikazano rješenje algoritma k-središnjih vrijednosti po varijablama.



Slika 27. Grafički prikaz algoritma k-središnjih vrijednosti (Izvor: ekran autora)

6. Zaključak

U današnje vrijeme kada se tehnologija, a posebice ona informacijska, sve više i više razvija dolazi do sve većeg prijenosa informacija. Same informacije odnosno podaci su beskorisni ako nisu svrstani u neki kontekst u kojem imaju neko značenje. S toga je vrlo važno grupirati podatke kako bi se lakše došlo do novih spoznaja. Bez grupiranja podataka za neka otkrića bilo bi potrebno uložiti mnogo više truda i resursa. Dok druga otkrića ne bi uopće bila moguća.

Ljudske djelatnosti poput medicine, sociologije, menadžmenta, istraživanje, i druge uopće nisu zamislive bez grupiranja podataka. U svim navedenim djelatnostima potrebno je grupirati podatke iz neke velike grupu u neku manju grupu kako bi im se odredile neke međusobne sličnosti te iz tih sličnosti zaključilo nešto novo.

Osim što se analizom i grupiranjem podataka lakše zaključuje o pojedinim objektima ti objekti svrstavaju se u pregledne skupine. Istraživačima vrlo je bitno da se koriste preglednim podacima kako ne bi dolazilo do neželjenih grešaka. Baš zbog toga analiza i obrada podataka i grupa jedan je od najvažnijih ljudskih procesa u modernom svijetu.

U ovom radu prikazan je stvarni primjer grupiranja podataka korištenjem algoritma k-središnjih vrijednosti u programu R. Kao što je prikazano u radu svaki od algoritama ima i prednosti i mane. Dobiveni rezultati su analizirani i vidljivo je kako u danom primjeru algoritam k-središnjih vrijednosti daje zadovoljavajuće rezultate, ali ne i u potpunosti točne.

Popis literature

Knjige i članci:

1. Devčić K., Tonković Pražić I., Župan Ž. (2012), Grupna analiza: primjena u marketinškim istraživanjima, Veleučilište Nikola Tesla u Gospiću
2. Hair, J.F. Jr., Black W.C., Babin B.J., Anderson R.E. (2010), Multivariate Dana Analysis, 7th Edition, Pearson Prentice Hall
3. Kiš Ž. (2012), Grupna analiza i njezina primjena u bankarstvu, diplomski rad, Ekonomski fakultet u Osijeku
4. Mirošević I. (2016), Algoritam k-sredina, stručni rad, KOG
5. Nachtwey A., Riedel R.,Mueller E. (2009), Cluster Analysis as a Method for the Planning of Production Systems , Computers & Industrial Engineering, 53:64
6. Nepoznati autor (2015), Multivarijantna statistička analiza, Predavanja iz kolegija , Ekonomski fakultet , Subotica
7. Sabo K., Scitovski R., Vazler I (2010)., Grupiranje podataka: klasteri, Osječki matematički list 10, 149:178

Internet izvori:

8. Academia Edu, dostupno na: http://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_ Mining, datum pristupa: 21.7.2020.
9. Ekonomski fakultet u Zagrebu, sažetak predavanja, 2020., dostupno na: http://www.ffzg.unizg.hr/psiho/phm/nastava/pmm/KLASTER_SAZETAK_PREDAVANJA_ZA_STUDENTE.DOC, datum pristupa: 21.7.2020.
10. IBM, dostupno na: https://www.ibm.com/developerworks/library/os-weka2/, datum pristupa: 21.7.2020.
11. Institut Ruđer Bošković (2020), Otkrivanje znanja dubinskom analizom podataka, Priručnik za istraživače i studente, dostupno na: http://lis.irb.hr/Prirucnik/prirucnikotkrivanje-znanja.pdf, datum pristupa: 21.7.2020.

Popis slika

Popis tablica