Skupina 22: *k*-means algoritem

Poročilo

Projekt v povezavi s predmetom Operacijske raziskave

Avtorja:
Matevž Raspet, Eva Šraj

Ljubljana, november 2018

Kazalo

Slike

1 KRATEK OPIS PROBLEMA

Pri projektu iz predmeta Operacijskih raziskav bova preučevala delovanje k-means algoritma.

NAVODILO: Zgeneriraj in predstavi naključne podatke v prostorih $\mathbb{R}$, $\mathbb{R}^2$ in $\mathbb{R}^3$, ki imajo strukturo množice in uporabi k-means algoritem s funkcijama oddaljenosti - **metoda Manhattan** in **metoda najmanjših kvadratov** (krajše MNK). Preizkusi k-means algoritem na nizu podatkov, ki imajo 10% bolj odstopajočih podatkov in primerjaj rezultate z uporabo obeh metod. Za podatke, zgenerirane v $\mathbb{R}^2$, poišči povezavo med k-means algoritmom z Vorionojevim diagramom in povezavo ponazori grafično.

1.1 Kaj je *k*-means algoritem in kako deluje?

To je metoda vektorske kvantizacije, namenjena za analizo grupiranja podatkov, predvsem pri rudarjenju s podatki. Ta algoritem naredi particijo n meritev v k različnih množic, v katerih vsaka meritev pripada množici z najbližjo povprečno vrednostjo. Rezultat particije je razdelitev prostora v Voronojeve celice.
BOLJ FORMALNO: Naj bodo $(x_1, x_2, \ldots, x_n)$ dane meritve, kjer je vsaka meritev x_i d-dimenzionalen realni vektor. K-means algoritem napravi particijo n meritev v $k(\leq n)$ množic $S = \{S_1, S_2, \ldots, S_k\}$, na podlagi najmanjše vsote kvadratov razdalj med meritvami posamezne množice in njihovim povprečjem μ_i.

Torej, s k-means algoritmom opravimo grupiranje n naključnih podatkov v množice v treh korakih:

- narišemo k naključno izbranih točk v prostoru (prvotne vrednosti povprečnih vrednosti) $\mu_1^{(1)}, \ldots, \mu_k^{(1)}$;
- v zadolžitvenem koraku (*angl. assignment step*), vsako točko določimo najbljižji povprečni vrednosti
  $S_i^{(t)} = \{x_j : \|x_j - \mu_i^{(t)}\|^2 \leq \|x_j - \mu_{i'}^{(t)}\|^2 \ \forall i' \in \{1, \ldots, k\}\}$
- v koraku nadgrajevanja (*angl. upgrade step*), zgeneriramo nove centroide množic

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Problem je sicer računsko zahteven, ker se ga ne da rešiti v polinomskem času (**NP-zahteven** algoritem). Vendar pa obstajajo učinkoviti hevristični

algoritmi, ki se pogosto uporabljajo in se hitro približajo lokalnemu optimumu.

Z najinim projektom sva se osredotočila predvsem na delovanje k-means algoritma in ga preizkusila na konkretnih podatkih v prostorih $\mathbb{R}$, $\mathbb{R}^2$ in $\mathbb{R}^3$.

1.2 Neznani pojmi

Kot sva omenila že zgoraj, sva uporabila 2 metodi za računanje razdalj - **metoda Manhattan** in **metoda najmanjših kvadratov** (krajše MNK).

Definicija: Naj bosta X in Y n-razsežna vektorja; $X = (x_1, x_2, \ldots, x_n)$ in $Y = (y_1, y_2, \ldots, y_n)$.
Razdalja po **metodi Manhattan** je definirana kot:

$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

Po **metodi najmanjših kvadratov** razdaljo definiramo kot:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

Voronojev diagram

V matematiki Voronojev diagram prikazuje razdelitev ploskve na območja, osnovana na najmanjši razdalji točk do centra v posameznih množicah. Obarva nam podatke, ki spadajo k istemu centru.

2 PODATKI V PROSTORU $\mathbb{R}$

CILJ: S pomočjo k-means algoritma sva na podlagi deleža posameznih krvnih skupin v izbranih evropskih državah določila tipe krvnih skupin.

Iz Wikipedije sva v program RStudio uvozila podatke o porazdelitvi krvnih skupin po državah. Zaradi boljše preglednosti na grafu in boljšega ujemanja podatkov sva se omejila le na evropske države.

PODATKI:
Za 37 izbranih evropskih držav sva uvozila deleže posameznih krvnih skupin v tabelo. Za vsako državo sva pogledala porazdelitev na 8 krvnih skupin, imenovanih 0+, A+, B+, AB+, 0-, A-, B- in AB-.

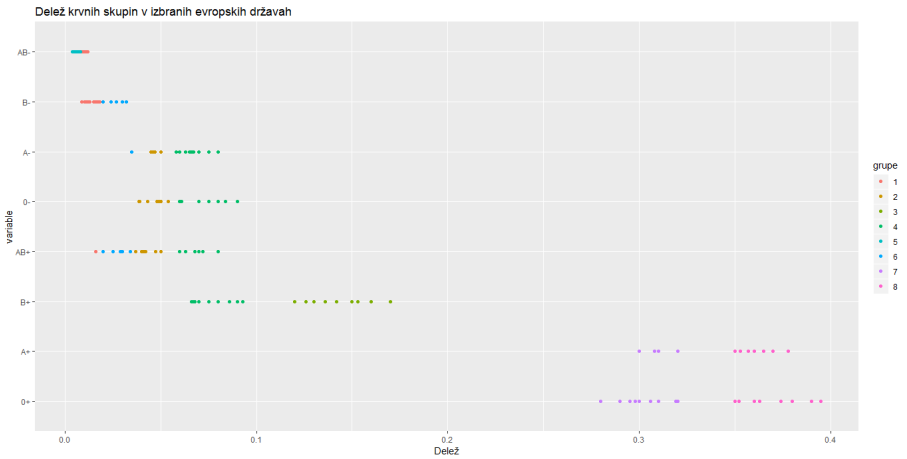
Glede na formalno definicijo k-means algoritma so podatki:

- $x_1, \ldots, x_{37}$, kjer x_i prikazuje izbrano evropsko državo,
- $\forall x_i (i = 1, \ldots, 37)$ je 8-dimanzionalni vektor $\Rightarrow$ v vsaki državi je 8 tipov krvnih skupin

S k-means algoritmom sva napravila particijo 37 meritev v 8 množic,
$S = \{S_1, \ldots, S_8\}$.
Množice nisva poimenovala, ampak samo oštevilčila od 1 do 8.

REZULTATI:

Slika 1: Graf MNK: Delež krvnih skupin v izbranih evropskih državah



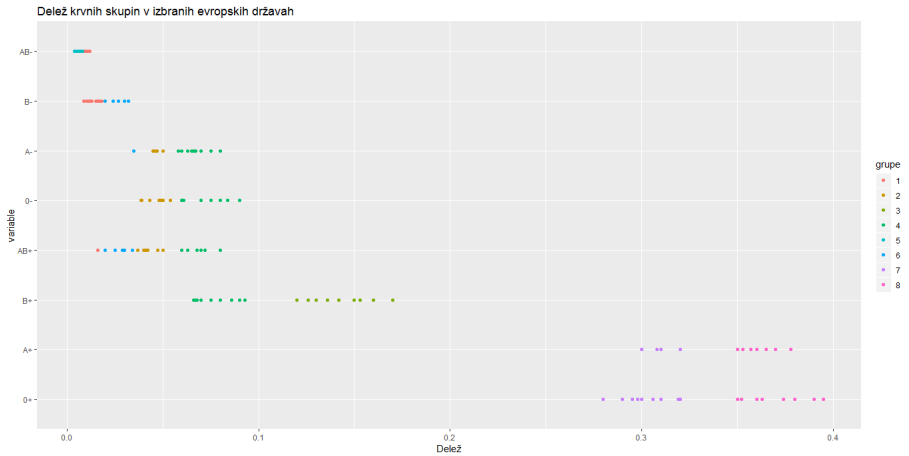
Kot je razvidno iz grafa zgoraj, se število krvnih skupin znotraj iste množice razlikuje.
V množici S_1 imamo samo krvno skupino AB-, medtem ko so v množici S_5 kar 4 (A-, 0-, AB+, B+). Po drugi strani pa ima množica S_6 samo krvno skupino B+, kar pomeni da sva v tej množici uspela pogrupirati samo podatke iz ene krvne skupine, B+.

Poskušala sva pogrupirati podatke na način, da bi bila v vsaki množici natanko ena krvna skupina, kar pa nama je uspelo le na množicah S_1 in S_6. V vseh ostalih pa so zastopani tipi večih različnih krvnih skupin, kar pa ni optimalno.

Deleži krvnih skupin so med evropskimi državami podobno zastopane – najmanj ljudi ima krvno skupino AB-, najpogosteje pa sta zastopani krvni skupini 0+ in A+. Bolj gosto kot so narisane pikice na grafu, več držav ima podoben delež iste krvne skupne, takšna je krvna skupina AB-. Po drugi strani pa bolj poredko narisane pikice predstavljajo večja odstopanja deležev med državami znotraj posamezne krvne skupine. Kar je najbolj očitno pri krvni skupini 0+ (na Madržarskem in Češkem je delež le-te samo 0.27, v Islandija pa ta znaša 0.476).

Glede na vse zgoraj opisano, je možno podatke še malo prečistiti. Tako sva znotraj vsake množice z MNK poračunala oddaljenost posameznega podatka do centra. Te razdalje imajo vrednosti med $5.714 \cdot 10^{-6}$ (Velika Britanija, AB+) in $1.053 \cdot 10^{-1}$ (Irska, 0+). Ker vrednosti bolj ali manj odstopajo od centra, sva 10% najbolj odstopajočih izvzela.

Slika 2: Graf Manhattan: Delež krvnih skupin v izbranih evropskih državah



Zgornji graf predstavlja delež krvnih skupin evropskih držav, kjer sva izvzela

29 najbolj odstopajočih podatkov. Najbolj očitna razlika s prejšnjim grafom je zastopanost krvnih skupin A+ in 0+. To pa je povsem razumljivo, saj so pri teh 2 krvnih skupinah podatki najbolj razpršeni.

Poleg tega sva znotraj vsake množice poračunala tudi skupno vsoto razdalj podatkov do centrov, dobljeno po MNK. Razdalje pri množicah S_1 (temno modra barva), S_2 (rdeča barva), S_5 (temno zelena barva) in S_6 (svetlo zelena barva) so tudi po odstranitvi 10% najbolj odstopajočih meritev ostale enake. Kar pomeni, da podatki od prej navedenih množic niso najbolj odstopali od centra. Medtem ko so se pri ostalih 4 množicah skupne razdalje znotraj iste množice očitno zmanjšale.

Poleg metode MNK sva za računanje razdalj uporabila tudi metodo Manhattan. Poleg tega, sva narisala tudi grafa [graf.Manhattan] in [graf.Manhattan.outliers] (v datoteki prostorRkri.R), ki pa sta nama dala enake vrednosti in rezultate kot pri metodi MNK.

TEORETIČNA UTEMELJITEV: V enačbah (1) in (2) sta neznanki:
x_i..... delež krvne skupine za vsako izbrano evropsko državo
y_i..... center posamezne množice, definirane s k-means algoritmom

$$|x_i - y_i| = \sqrt{(x_i - y_i)^2}$$

3 PODATKI V PROSTORU $\mathbb{R}^2$

3.1 PRIMER 1

CILJ: S pomočjo k-means algoritma sva poskušala na podlagi statistike igralcev (zadeti goli in podeljene asistence), določiti igralne pozicije igralcev.

Iz spleta sva uvozila podatke o 100 svetovno znanih nogometaših.

PODATKI:
Za vsakega igralca naju je zanimalo število zadetih golov in asistenc in na podlagi teh podatkov sva poskušala določiti igralne pozicije.

Za lažjo predstavo sva najprej narisala [graf1] (v datoteki prostorR2goli.R), ki prikazuje dejanske pozicije glede na statistiko.
Nato sva s pomočjo podatkov naredila grupiranje v tri skupine. Podatke sva poskušala pogrupirati glede na igralne pozicije tipov:

- napad: centre forward, second striker,
- sredina: attacking midfield, central midfield, left winger, right winger,
- obramba: centre back, defensive midfield, goalkeeper, right back, left back.

Glede na formalno definicijo k-means algoritma so podatki:

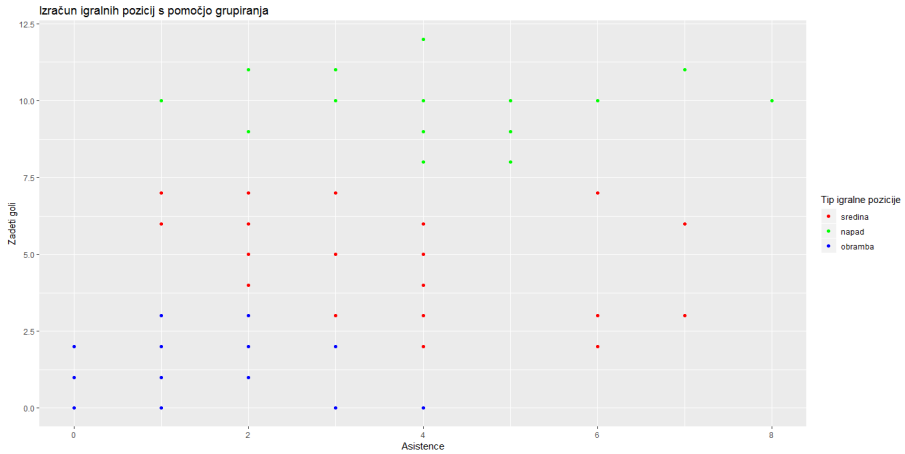
- $x_1, \ldots, x_{100}$, kjer x_i prikazuje posameznega nogometaša,
- $\forall x_i (i = 1, \ldots, 100)$ je 2-dimanzionalni vektor $\Rightarrow$ prva komponenta prikazuje število asistenc posameznega igralca, druga komponenta pa število zadetih golov.

Z algoritmom sva napravila particijo 100 meritev v 3 množice, $S = \{S_1, S_2, S_3\}$, kjer S_1 predstavlja obrambo, S_2 sredino in S_3 napad.

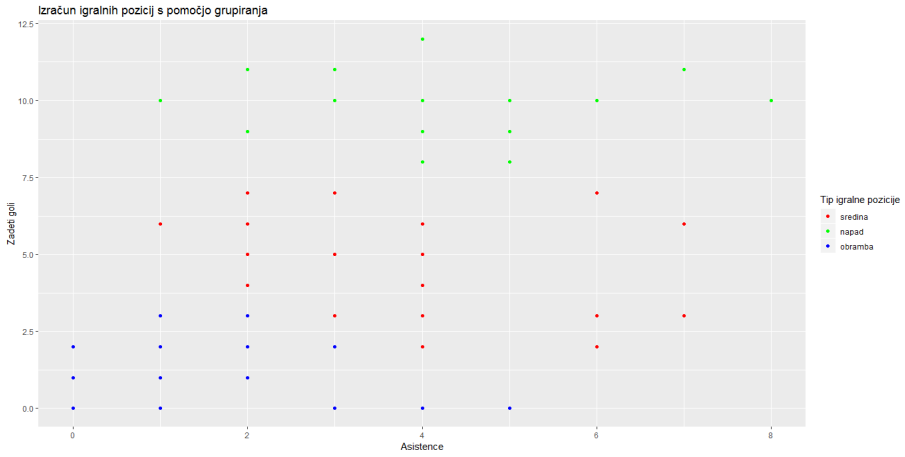
REZULTATI:

Iz dobljenih rezultatov in dejanskih igralnih pozicij lahko vidimo, da je k-means algoritem večinoma pravilno pogrupiral podatke v množice napad, sredina in obramba.
V množici napad (modra barva) so najbolj zastopani podatki s pozicijami klasičnih napadalcev(*angl.centre forward*) in pa polnapadalec (*angl. second striker*), v množici sredina (zelena barva) igralci tipa centralnih veznih igralcev(*angl. central midfield*), v množici obramba (rdeča barva) pa vratarji (angl. goalkeeper) in pa levi ter desni bočni igralci.

Slika 3: Graf MNK: Izračun igralnih pozicij s pomočjo grupiranja



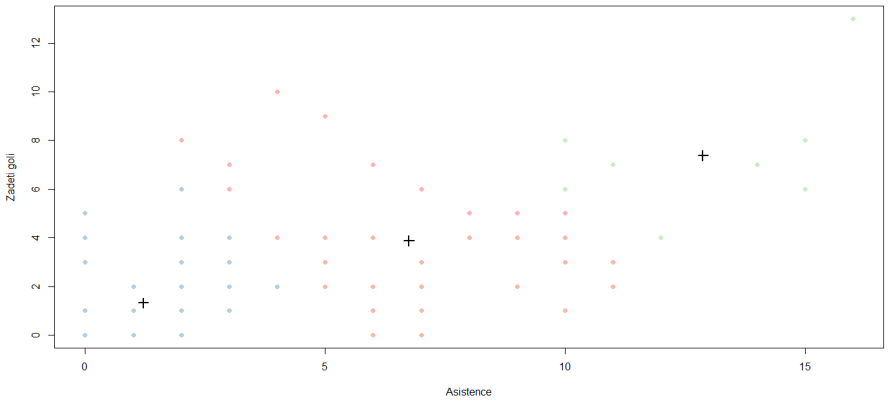
Slika 4: Graf MNK: Izračun igralnih pozicij s pomočjo grupiranja



Prav tako sva s k-means algoritmom dobila povprečno število zadetih golov in asistenc po tipih igralnih pozicij napad, sredina in obramba.
Nekateri podatki so bili razvrščeni v napačne razrede, kar je bila posledica različnih dejavnikov. Predvsem so to napadalni igralci, ki niso izplonili svojih statistik zaradi poškodbe (niso zadevali, saj niso igrali vseh tekem), prav tako pa se je pojavil kakšen obrambni igralec (vendar redko), ki je izstopal iz svoje množice, saj je zadel več golov oziroma podelil več asistenc, kot se za igralca obrambne vrste pričakuje.

Prav tak opa sva za primer v $\mathbb{R}^2$ naredila Voronojev diagram.
Ta nama je razdelil območje na 3 različne dele.

Slika 5: Voronojev diagram



3.2 PRIMER 2

CILJ: Z algoritmom sva poskušala določiti kritična leta umrljivosti v Sloveniji v letu 2017.

Podatki, ki sva jih na spletu našla v obliki .csv, prikazujejo število umrlih ljudi po starostih.

PODATKI:
Za vsako starost posebej (od 0 pa do 100+ let) sva pogledala število umrlih moških in žensk.
V tabeli in na grafu sva zaradi večje preglednosti moške označila z 1 in ženske z 2.

Na začetku sva narisala [graf2] (v datoteki prostorR2umrli.R), kjer sva pogledla pogostost smrti za vsako starost, ločeno glede na spol. Nato sva s pomočjo k-means algoritma vse podatke pogrupirala v 3 množice glede na kritično starost. Te so:

- najbolj - rdeča barva na grafu;
- srednje - modra barva na grafu;
- najmanj - zelena barva na grafu.

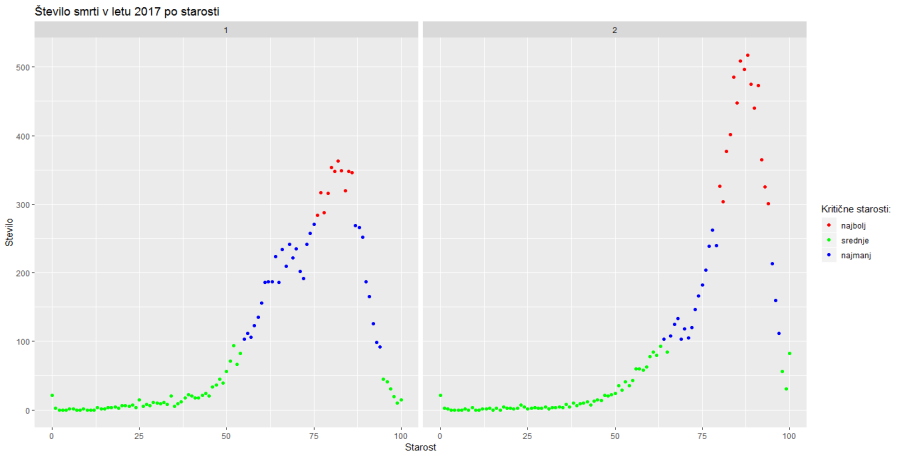
Glede na formalno definicijo k-means algoritma so podatki:

- $\forall x_i (i = 1, \ldots, 202)$ je 2-dimanzionalni vektor $\Rightarrow$ Prva komponenta vektorja je starost, druga pa število umrlih.

Napravila sva particijo 202 meritev v 3 množice, $S = \{S_1, S_2, S_3\}$, kjer S_1 predstavlja najbolj kritična leta smrti, S_2 srednje in S_3 najmanj kritična leta.

REZULTATI:

Slika 6: Graf MNK: Število smrti v letu 2017 po starosti

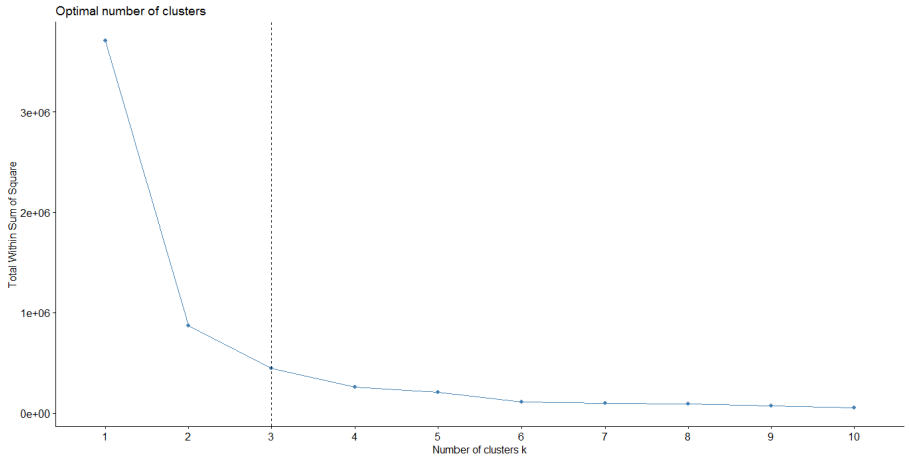


Glede na dane rezultate, ki jih vrne k-means algoritem, lahko vidimo, da nam je večino podatkov pogrupiral v skladu s pričakovano krivuljo umrljivosti po starosti. Na grafu najbolj kritično starost predstavlja rdeča barva. To so moški med 76 in 86 letom ter ženske med leti 80 in 94. Rezultat ni presenetljiv, saj imajo ženske daljšo življenjsko dobo, kar prikazuje tudi najvišja višina pikic bolj desno na grafu 2. Z algoritmom pa sva najmanj optimalno dobila množico S_2 (zelena barva). Le-ta vsebuje tako novorojenčke kot tudi visoke starostnike. Ker je tu razpon med leti največji, je tudi vsota razdalj do centra največja.

Prav tako pa sva tudi za ta primer naredil Voronojev diagram, ki nama je razdelil območje na 3 dele.

Na koncu sva na tem primeru z Elbowo metodo sprogramirala iskanje oprimalnega k in dobila naslednji graf.

Slika 7: Optimalen k



4 PODATKI V PROSTORU $\mathbb{R}^3$

CILJ: Poskušala sva pogrupirati avtomobile v skupine na podlagi izbranih karakteristik.

Podatke sva uvozila s pomočjo že vgrajene tabele mtcars.
Na podlagi dobljenih rezultatov se lahko bralec odloči, katere znamke avtomobilov mu bolj ustrezajo, glede na preučevane karakteristike.

PODATKI:
Za vsak avtomobil so naju zanimali podatki glede na 3 karakteristike:

- število prevoženih kilometrov z enim galonom goriva;
- število konjskih moči;
- najmanjši izmerjeni čas na četrtino milje.

Glede na formalno definicijo k-means algoritma so podatki:

- $x_1, \ldots, x_{32}$, kjer x_i prikazuje posamezen model avtomobila,
- $\forall x_i (i = 1, \ldots, 32)$ je 3-dimanzionalni vektor $\Rightarrow$ za vsak avto so naju zanimale 3 (prej navedene) karakteristike

S k-means algoritmom sva napravila particijo 32 meritev v 4 množice, $S = \{S_1, \ldots, S_4\}$. Množice nisva poimenovala, ampak samo različno obarvala (zelena, modra, rumena, vijolična).

REZULTATI:
Rezultati so bili podani v štirih različnih množicah.
Množica S_1, ki je zelene barve, predstavlja znamke avtomobilov z najmanjšo porabo, vendar zato tudi majhno število konjskih moči in pa posledično daljši čas na četrtino milje.
Množica S_2, obarvana z modro, predstavlja optimalne izbire avtomobilskih znamk za vsakdanje voznike. Gre za tipe avtomobilov, ki lahko prevozijo do 30 milj z eno galono goriva, prav tako pa imajo zadosti konjskih moči (čez 100 hp) in sorazmerno nizek čas za prevoz četrtine milje (17 sekund).
V tretji množici S_3, obarvani z rumeno, imajo podatki najmanjšo vsoto razdalj do njihovega centra (najbolj prilegajoči se podatki centru). Le-ta vsebuje avtomobilske znamke, ki imajo dokaj podobno porabo, kot avtomobili iz zadnje množice (vijolično obarvani). Vendar imajo avti v S_3 precej manj konjskih moči in daljši čas za prevoz četrtine milje.
Tako bi se kot kupec s stališča porabe, pri izbiri med množicama S_3 in S_4, raje odločil za nakup avtomobila iz množice S_4.

5 VIRI

L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, John Wiley Sons, Chichester, UK, 2005.

J. Kogan, *Introduction to Clustering Large and High-dimensional Data*, Cambridge University Press, New York, 2007.