Univerzitet u Novom Sadu
Prirodno-matematički fakultet
Departman za matematiku i informatiku

Analiza i statistička obrada podataka *FitBit* pametnih uređaja

Seminarski rad

Marina Adamović 100m/23

Septembar, 2024

Sadržaj

1 Uvod

FitBit je kompanija koja nudi širok spektar pametnih bežičnih uređaja za praćenje fizičkih aktivnosti i promene u telu usled istih. Zahvaljujući finoj tehnologiji i osetljivim senzorima, ovi uređaji omogućavaju precizno merenje otkucaja srca, broja pređenih koraka, kalorijskog utroška kao i opšteg kvaliteta sna, nivoa aktivnosti i mnogih drugih parametara. U ovom radu obrađuje se skup podataka dobijen beleženjem aktivnosti putem FitBit pametnih satova od strane trideset tri korisnika u periodu od 12. 4. 2016. do 12. 5. 2016. (izvor: FitBit skup podataka), sa ciljem sticanja uvida u samu strukturu korisnika i uviđanja interesantnih veza među beleženim parametrima koji mogu indukovati potencijalno korisne predloge za poboljšanje aplikacije i marketinških strategija ove kompanije. S obzirom na prirodu ovog skupa podataka, njegovom analizom može se doći do opšte koristih znanja o (ne)zdravim šablonima u svakodnevnom životu čoveka.

2 Analiza

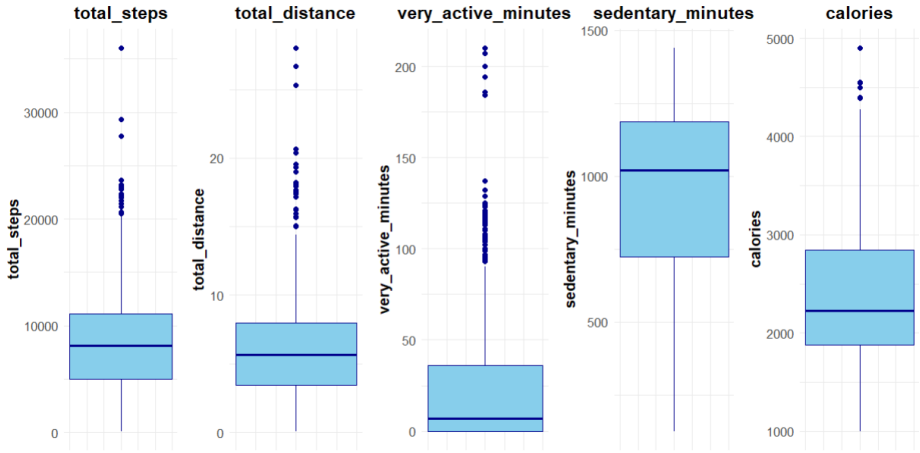
Gorenavedeni izvor sadrži čak dvadeset osam tabela koje pokrivaju navedeni period, kao i trideset dana pre njega. Međutim, kako se one u širokom opsegu razlikuju u svojoj granularnosti, dostupnosti i generalnom kvalitetu, fokus će biti na četiri najkompletnije tabele koje sadrže podatke o dnevnom nivou aktivnosti, kalorijskom utrošku i trajanju sna, kao i broju pređenih koraka i utrošenih kalorija po satu. Ove tabele se mogu pronaći u odeljku pod odgovarajućim vremenskim opsegom i to pod nazivima dailyActivity_merged (preimenovana u dailyActivity za potrebe rada), sleepDay_merged (dailySleep), hourlyCalories_merged (hourlyCalories) i hourlySteps_merged (hourlySteps).
Kategorisati ljude po ovim parametrima bez prethodnog znanja o njihovom polu, uzrastu ili fizičkim sklonostima nije jednostavan zadatak, te za potrebe ovog rada pretpostavićemo da je ovaj uzorak reprezentativan, odnosno da proizvedeni zaključci jesu primenjivi na prosečnog korisnika ovog uređaja.
Prilikom učitavanja navedenih tabela odbacujemo kolone koje nećemo posmatrati, zadržavamo one koje će nam korisiti, standardizujemo njihova imena i formatiramo ih na odgovarajući način. Takođe, brišemo duplirane unose, zapise sa nedostajućim podacima, a zatim i one profile koji su u tabelama imali znatno manji broj zapisa od očekivanog, te nad tako očišćenim skupom možemo započeti analizu. Za analizu korišćen je programski jezik *R*.

2.1 Tabela 1: dailyActivity

	id	date	total_steps	total_distance	very_active_minutes	sedentary_minutes	calories
1	1503960366	2016-04-12	13162	8.50	25	728	1985
2	1503960366	2016-04-13	10735	6.97	21	776	1797
3	1503960366	2016-04-14	10460	6.74	30	1218	1776
4	1503960366	2016-04-15	9762	6.28	29	726	1745
5	1503960366	2016-04-16	12669	8.16	36	773	1863

Slika 1: Zaglavlje tabele dailyActivity

Kolone *very_active_minutes* i *sedentary_minutes* predstavljaju broj minuta provedenih u visoko intenzivnom fizičkom naporu, odnosno mirovanju, respektivno. Kolona *total_distance* data je u km, dok se ostali atributi razumeju intuitivno. Izopštićemo zapise koji broje manje od 100 koraka i 1000 utrošenih kalorija, kako se takvi mogu smatrati nerelevantnim. Ova tabela nakon ulazne obrade broji 847 redova, među kojima je 32 korisnika.
Osnovne karakteristike numeričkih atributa predstavljene su u vidu box plot-a, na kom možemo videti važne indikatore raspodele podataka poput kvartila, medijane i outlier-a. Na primer, ovde možemo videti da većinu uzorka čine ljudi koji nemaju fizički zahtevan posao, budući da se fizičke aktivnosti preko 75 minuta smatraju outlier-ima, a u stanju mirovanja ispitanici provode u proseku 16.5 sati dnevno. Takođe, zanimljiv podatak je da ispitanici u proseku prepešače oko 5km u toku dana.



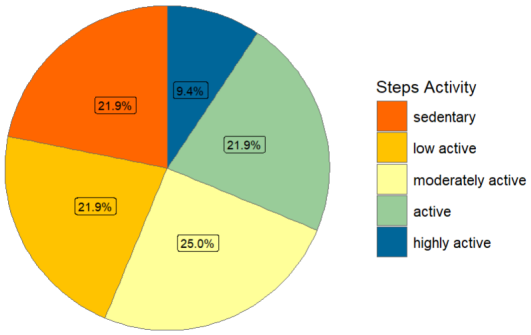
Slika 2: dailyActivity box plot

Kategorisaćemo korisnike na osnovu prosečnog broja pređenih koraka u toku dana na sledeći način: (izvor)

- Sedentary: manje od 5000 koraka
- Low active: od 5000 do 7500 koraka

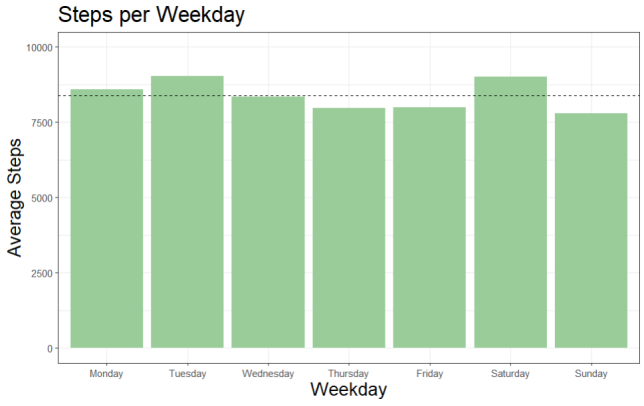
- Moderately active: od 7500 do 10000 koraka
- Active: od 10000 do 12500 koraka
- Highly active: više od 12500 koraka

Na grafiku ispod možemo videti da 21.9% uzoračkog skupa čine ljudi koji u proseku naprave manje od 5000 koraka dnevno, dok 31.3% njih premaši preporučenih 10000 koraka.



Slika 3: Distribucija korisnika po prosečnom broju koraka

Odgovor na pitanje - da li se svakim danom u nedelji pređe isti broj koraka - mogao bi biti interesantan moderatorima aplikacije iz više razloga. U tu svrhu, podatke grupišemo po datumima, svrstavamo ih u dane u nedelji i pokrećemo hi-kvadrat test fitovanja koji inicijalno pretpostavlja da se broj koraka ne razlikuje značajno među danima.



Slika 4: Prosečan broj koraka po danima u nedelji

Zaista, na stubičastom dijagramu 4 čini se da su razlike minimalne i ne tako značajne među stubovima. Međutim, hi-kvadratni test daje sledeće rezultate:

```
        Pearson's Chi-squared test

data:  cont_table
X-squared = 90.21, df = 6, p-value < 2.2e-16
```

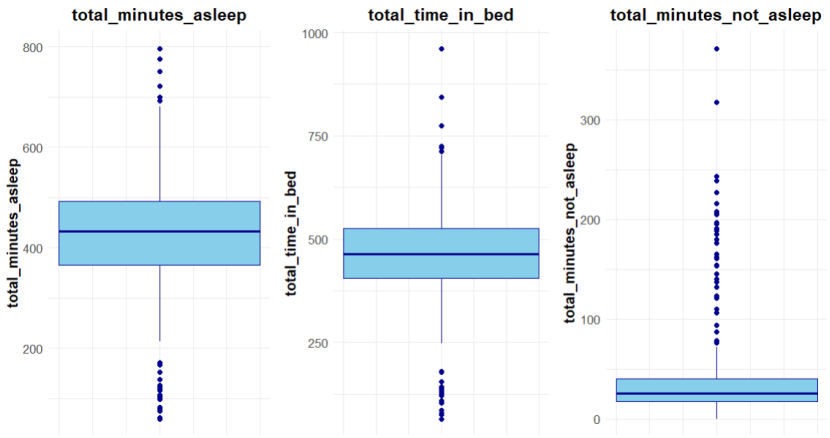
Niska p-vrednost ukazuje na odbacivanje nulte hipoteze i prihvatanje alternativne, koja kaže da se razlika među danima nije desila slučajno. Nedelja je rangirana kao najmanje aktivni, a utorak i subota kao najaktivniji dani. Jedna ideja za korišćenje ove informacije radi poboljšanja aplikacije bila bi uvođenje opcije za distinktivno postavljanje cilja u broju koraka za različite dane u nedelji, ili čak isključivanje istog u potpunosti, kako se ljudi ne bi osećali loše ukoliko, recimo, nedeljom ne isunjavaju svoj cilj i time gubili motivaciju za korišćenjem uređaja.

2.2 Tabela 2: dailySleep

	id	date	total_minutes_asleep	total_time_in_bed	total_minutes_not_asleep
1	1503960366	2016-04-12	327	346	19
2	1503960366	2016-04-13	384	407	23
3	1503960366	2016-04-15	412	442	30
4	1503960366	2016-04-16	340	367	27
5	1503960366	2016-04-17	700	712	12

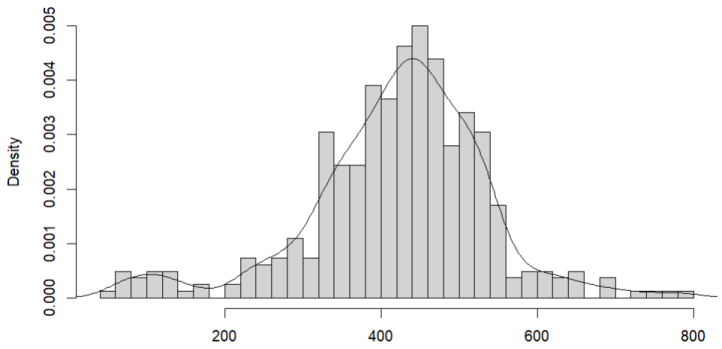
Slika 5: Zaglavlje tabele dailySleep

Kolona *total_minutes_not_asleep* predstavlja vreme koje je bilo potrebno da korisnik zaspi, kao i minute tokom kojih nije spavao tokom sesije snimanja. Ova kolona može da bude dobar pokazatelj kvaliteta sna, gde niže vrednosti predstavljaju kvalitetniji san, kako zbog kraćeg vremena potrebnog da se zaspi, tako i zbog veće konzistentnosti sna. Ova tabela broji 407 redova, među kojima je 22 ispitanika.



Slika 6: dailySleep box plot

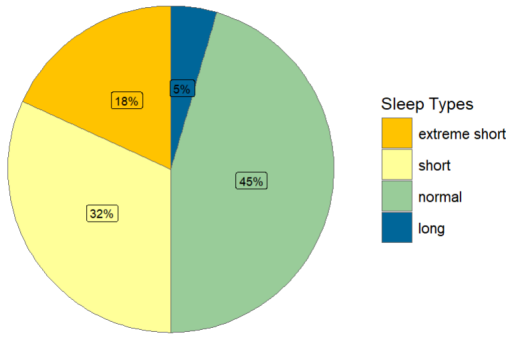
Sa Slike 6 možemo zaključiti da su brojni outlier-i prisutni u zabeleženim sesijama spavanja. Takvi podaci mogu biti rezultati tehničkih grešaka prilikom snimanja, ali i realni. Kako ne postoji opšte pravilo uslovljavanja kojim možemo utvrditi uzrok pojave odstupanja, posmatramo Sliku 7.



Slika 7: Histogram zapisa spavanja

Na ovom histogramu vidimo da vreme spavanja prati približno normalnu raspodelu, što je očekivano nad ovom količinom zapisa. Međutim, primećujemo da su izuzeci dužih sesija manjebrojni u poređenju sa izuzecima kraćih sesija, koje narušavaju normalnost raspodele. Nakon filtriranja kraćih sesija i grupisanja istih po ID-u, saznajemo da je među njima prisutno čak 10 različitih ljudi, što je skoro polovina posmatranih ispitanika. Odlučujemo da ih zadržimo za sada kao realne podatke, kako ne možemo zasigurno pripisati neispravnom upravljanju niti pojedincima.
U nastavku kategorišemo ispitanike u četiri grupe na osnovu njihovog prosečnog trajanja spavanja:

- Extreme short - do 5 sati sna
- Short - između 5 i 7 sati sna
- Normal - između 7 i 10 sati sna
- Long - preko 10 sati sna



Slika 8: Distribucija korisnika po proseku trajanja spavanja

Ova kategorizacija daje nam nekoliko korisnih informacija: čak 18% ispitanika spava u proseku manje od 5 sati, što može sa jedne strane ukazivati na slučaj neispravnog rukovanja ili pak na ozbiljni problem sa spavanjem poput *sleep apnea-e* i drugih disruptivnih šablona spavanja. U oba slučaja, jedna ideja za poboljšanje aplikacije bila bi implementacija linkova ka stručnoj literaturi koja sadrži savete i korisne informacije o ovakvim problemima, u onim slučajevima kada se postigne ovako nizak rezultat na kraju jednog ciklusa snimanja. Drugi ekstrem (dugo spavanje) može ukazivati na probleme poput depresije, bolesti srca i drugih zdravstvenih poteškoća, te je važno da korisnik takođe bude obavešten ukoliko prevazilazi gornju granicu normale. U našem slučaju, 45% ljudi spava preporučeno dugo, 5% predugo, dok druga polovina spava kratko. Dakle, 23% (5 ispitanika) u posmatranoj grupi bilo bi obavešteno da provere postavke uređaja i utvrde da rukuju njime valjano pre nego što nastave monitoring, i nakon toga, ukoliko su sve tehničke prepreke uklonjene, bili bi obavešteni i savetovani o potencijalnim zdravstvenim problemima.
Sada bismo želeli da proverimo da li preporučeni prosečan broj sati sna takođe znači i manji broj budnih minuta, odnosno kvalitetniji san. To ćemo utvrditi pomoću t-testa koji pokrećemo nad dve grupe ispitanika: one koji spavaju u proseku bar 7 sati dnevno, i one koji spavaju kraće od toga. Početna pretpostavka testa je da ne postoji statistički značajna razlika između ove grupe (iako se mi nadamo suprotnom). Rezultati testa:

```
        Welch Two Sample t-test

data:  under_7h$average_awake and over_7h$average_awake
t = -0.6143, df = 15.219, p-value = 0.5481
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -79.56530  43.92893
sample estimates:
mean of x mean of y
 36.90909  54.72727
```

Kako p-vrednost nije manja od 0.05, zadržavamo nultu hipotezu, odnosno zaključujemo da kvalitet sna po ovoj metrici ne zavisi od toga da li u proseku spavamo 7 sati ili ne.

2.3 Tabela 3: pridružene tabele dailyActivity i dailySleep

Nakon pregleda Tabele 1 i Tabele 2 iz prethodnih poglavlja, spajamo iste po ID-ju i datumu kako bismo iskombinovali kolone iz obe tabele i došli do novih zaključaka koji povezuju spavanje sa aktivnošću. Videli smo da konzistentnost korisnika nije bila uniformna kroz ove dve tabele, te možemo zaključiti da će u kombinovanoj tabeli neki redovi biti izgubljeni. Zaista, sada raspolažemo sa 403 zapisa.
Za početak, želimo da proverimo povezanost između vremena koje je potrebno da ispitanik zaspi i nivoa aktivnosti definisanog preko broja pređenih koraka. Pitamo se da li će ljudi koji prepešače bar 10000 dnevno imati manje prosečno vreme koje je potrebno da zaspe. Pokrećemo t-test koji inicijalno pretpostavlja da će srednje vrednosti ove dve grupe biti jednake, odnosno da se razlika desila slučajno.

```
        Welch Two Sample t-test

data:  under_10k$total_minutes_not_asleep and over_10k$total_minutes_not_asleep
t = -1.299, df = 301.42, p-value = 0.1949
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.157729   3.308276
sample estimates:
mean of x mean of y
 36.98739  43.41212
```

Rezultati testa daju p-vrednost 0.1949 koja je veća od 0.05, što nam govori da treba da prihvatimo nultu hipotezu koja kaže da je razlika srednjih vrednosti prosečnog vremena ove dve grupe zanemarljivo mala. Dakle, naša ideja o uticanju broja koraka na ovaj indikator kvaliteta spavanja nije proizvela koristan zaključak, što potencijalno može značiti da je to aktivnost nedovoljnog intenziteta da bi uticala dovoljno snažno, te se sada odlučujemo na posmatranje uticaja aktivnosti višeg intenziteta. Pitamo se da li ispitanici koji u proseku provedu određeni broj minuta u fizičkoj aktivnosti višeg intenziteta imaju manje prosečno vreme potrebno da zaspe (odnosno manje minuta bez sna tokom noći). Pokrenuli smo t-test na isti način kao malopre za nekoliko vrednosti i ispostavilo se da do 30 minuta nema primetnog uticaja, do 45 minuta p-vrednost

graduativno raste, a prva zaokružena minutaža za koju padne ispod 0.05 je 45 minuta.

```
        Welch Two Sample t-test

data:  over_45min$total_minutes_not_asleep and under_45min$total_minutes_not_asleep
t = -3.0964, df = 226.88, p-value = 0.002206
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.18164  -4.92893
sample estimates:
mean of x mean of y
 29.05618  42.61146
```

Ovo nas dovodi do zaključka da 45 minuta fizičke aktivnosti (većeg intenziteta od šetanja) u toku dana može poboljšati kvalitet našeg sna, dajući u proseku 13.5 minuta bez sna manje, odnosno minuta potrebnih da zaspimo.
Sledeće što želimo da saznamo je da li će minimalan poželjan broj koraka (5000) uticati na broj prespavanih sati tog dana, odnosno da li postoji povezanost ove niskointenzivne aktivnosti sa samim trajanjem spavanja. Ovaj broj smo birali na osnovu podele analogne onoj kada smo kategorisali korisnike: 5000 pređenih koraka smatra se granicom između sedenternog i aktivnog dana. Koristimo Fišerov test, gde nulta hipoteza kaže da statistički značajna povezanost ne postoji, a rezultati testa glase:

```
        Fisher's Exact Test for Count Data

data:  cont_table
p-value = 0.002579
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2683550 0.7804676
sample estimates:
odds ratio
 0.4623742
```

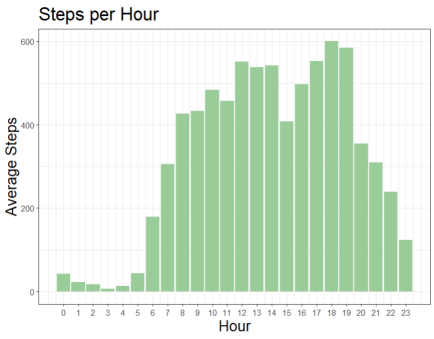
Dakle, kako je p-vrednost niža od 0.05, odbacujemo nultu hipotezu i zaključujemo da postoji povezanost između ovih grupa. Informacija o ovom odnosu može pomoći prilikom lociranja nezdravih šablona, na primer, kod ljudi koji imaju probleme sa spavanjem kod kojih ni ova granica aktivnosti ne indukuje pozitivnu promenu.

2.4 Tabela 4: pridružene tabele hourlyCalories i hourlySteps

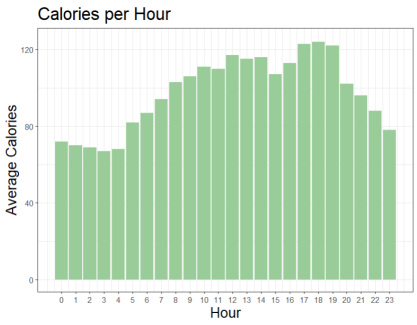
	id	activity_hour	calories	step_total
1	1503960366	2016-04-12 00:00:00	81	373
2	1503960366	2016-04-12 01:00:00	61	160
3	1503960366	2016-04-12 02:00:00	59	151
4	1503960366	2016-04-12 03:00:00	47	0
5	1503960366	2016-04-12 04:00:00	48	0

Slika 9: Zaglavlje tabele merged_hourly

Na Slici 9 prikazana je tabela koja odmah kombinuje tabele hourlyCalories i hourlySteps. U njima nema nepoklapanja što se tiče konzistentnosti beleženja, te ovim spajanjem ne gubimo zapise i konačna tabela broji 22011 redova, među kojima je 32 korisnika.
Na stubičastim grafikonima ispod prikazane su raspodele prosečnog broja koraka i prosečnog utroška kalorija po satima, respektivno. U 3 ujutru naši ispitanici su najmanje aktivni, dok se najviša aktivnost beleži u 18 časova. Sudeći po naglim skokovima u koracima na prvom dijagramu, ispostavlja se da po jedna trećina ljudi ustaje između 5 i 6, između 6 i 7, i između 7 i 8 časova ujutru. Sličan graduativni pad javlja se u večernjim časovima u prozorima od po jedan čas, počevši od 20 do 23 časa, što je takođe očekivani trend. Najveći pad javlja se između 19 i 20 časova.



Slika 10: Prosečan broj koraka po satima



Slika 11: Prosečan broj utrošenih kalorija po satima

Vizuelno se nazire sličnost u oblicima ove dve raspodele, te želimo da dublje ispitamo njihov odnos. Jasno je da šetanjem trošimo kalorije i izvesno je da će ove dve pojave prirodno biti proporcionalne. Međutim, kako čovek troši kalorije i u stanju apsolutnog mirovanja, a pogotovo prilikom neke fizičke aktivnosti, želeli bismo ugrubo da ispitamo kako konkretno broj koraka utiče na potrošnju kalorija

po satu. U tu svrhu svaki sat prikazujemo na tačkastom dijagramu pomoću ove dve numeričke karakteristike i konstruišemo linearni model po njima:

```
Call:
lm(formula = average_calories ~ average_steps, data = grouped_by_hour)

Coefficients:
  (Intercept)  average_steps
     68.66313        0.08956
```

Tumačenjem ovih koeficijenata dolazimo do sledećih zaključaka: jednom više pređenom koraku po satu u proseku odgovara dodatnih 0.08956 utrošenih kalorija po satu, te dolazimo do zaključka da moramo prešetati 11.16 koraka kako bismo utrošili 1 kaloriju. Presek sa y-osom, odnosno nula našeg linearnog modela je 68.66313. Drugim rečima, ako u toku jednog sata ne šetamo, očekuje se da ćemo svakako utrošiti 68.66 kalorija.
Produkt ovog modela, kao i prethodnih testova, ponovo, daju veoma grube aproksimacije koje svakako mogu postati preciznije uračunavanjem ostalih fizičkih karakteristika korisnika.

3 Zaključak

U ovom radu, kroz analizu priloženog seta podataka, došli smo do zaključaka koji omogućavaju bliži uvid u strukturu klijenata kompanije FitBit. Dato je nekoliko predloga za poboljšanje aplikacije u skladu sa rezultatima obavljenih testova i istaknute su neke opšte karakteristike u šablonima ponašanja korisnika koje se mogu dalje samostalno iskoristiti u marketingu i opštem poboljšanju usluge.
Svrha ovog seminarskog rada bila je ilustracija procesa analize podataka u programskom jeziku *R*.

4 Literatura

1. Analizirani skup podataka:
   https://www.kaggle.com/datasets/arashnic/fitbit

2. https://www.10000steps.org.au/learn-and-discover/counting-steps/#:~:text=Sedentary%20is%20less%20than%205%2C000,than%2010%2C000%20steps%20per%20day