Znanost o podacima: Postupci testiranja hipoteze

Bašković, Antonia

Sveučilište u Rijeci – Odjela za informatiku

Diplomski studij informatike – nastavnički smjer

Antonia Bašković

Znanost o podacima: Postupci testiranja hipoteze

Diplomski rad

Mentor: prof. dr. sc. Maja Matetić

Rijeka, rujan 2018.

Zadatak diplomskog rada

Statističko testiranje hipoteze se koristi kada donosimo odluke vezano uz populaciju samo na temelju uzorka informacije. Zadatak diplomskog rada je izvesti postupke statističkog testa na zadanom skupu podataka. Odabir postupka testiranja treba odabrati na temelju postavljene hipoteze koja se dokazuje kao i u ovisnosti o prirodi podataka koji mogu biti numerički i kategorički. Primjenjivat će se postupci test varijance ANOVA, Hi-kvadrat test nezavisnosti i Pearsonov korelacijski koeficijent.

Sažetak

Tema diplomskog rada su postupci testiranja hipoteza na određenom skupu podataka. Na početku ću objasniti što je znanost o podacima. Također ću objasniti kako se koristi testiranje hipoteze kada donosimo odluke vezane uz populaciju samo na temelju uzoraka informacije. Odabir postupaka testiranja izvršit ću na temelju postavljene hipoteze koja se dokazuje u ovisnosti o prirodi podataka koji mogu biti numerički i kategorički. U projektnom zadatku vršimo testiranje pomoću postupaka test varijance ANOVA, Hi-kvadrat test nezavisnosti i Pearsonov korelacijski koeficijent. Testiranje ću izvesti u programu RStudio. Dobivene rezultate ću usporediti i na temelju toga doći do zaključka.

Ključne riječi

- Znanost o podacima, dubinska analiza podataka, postupci testiranja hipoteze, test varijance ANOVA, Hi-kvadrat test nezavisnosti, Pearsonov korelacijski koeficijent, p-vrijednost

Sadržaj

1. Uvod

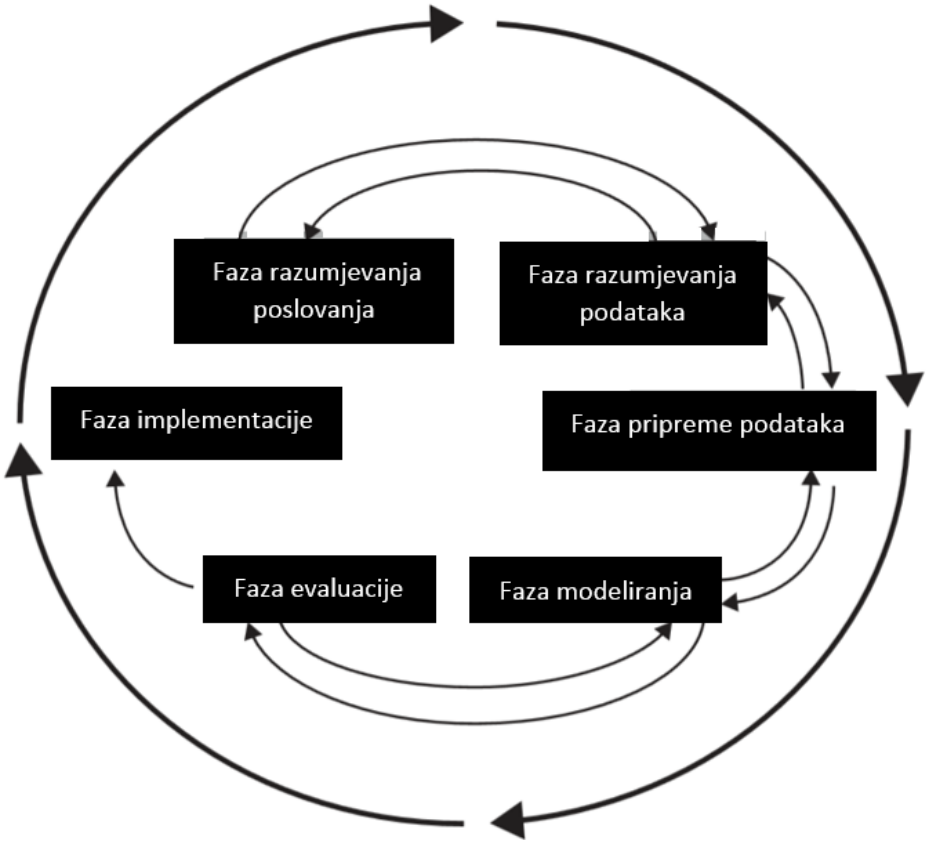
Znanost o podacima je interdisciplinarno polje koje se bavi otkrivanjem novih korelacija, uzoraka i trendova ispitujući veliku količinu podataka pohranjenih u baze koristeći tehnologije prepoznavanja uzoraka, kao i statističke i matematičke tehnike. Danas postoji niz pojmova koji se koriste za opisivanje ovog procesa, uključujući analitiku, prediktivnu analizu, „big data“, strojno učenje i otkrivanje znanja u bazama podataka, a svi oni dijele zajednički cilj pronalaska korisnog znanja iz velikih skupova podataka. Kroz ostatak ovog rada koristit ću pojam „data mining“ ili dubinska analiza podataka.

Jedna od grešaka povezanih s implementacijom dubinske analize podataka jest da se dubinska analiza podataka predstavlja kao izolirani skup alata, koji se primjenjuje u izoliranom postupku za analizu, koji je minimalno povezan s glavnim poslovnim ili istraživačkim naporima. Organizacije koje na taj način nastoje implementirati dubinsku analizu podataka imat će vidno smanjenu vjerojatnost uspjeha. Dubinsku analizu podataka treba promatrati kao proces.

Standardni proces koji se koristi jest CRISP-DM okvir: standardni proces za dubinsku analizu podataka. CRISP-DM zahtijeva da se dubinska analiza podataka smatra cjelovitim procesom, od komunikacije poslovnog problema, prikupljanja i upravljanja podacima, predprocesiranja podataka, stvaranja modela, procjene modela i implementacije modela.

1.1. CRISP-DM

Prema CRISP-DM određeni projekt dubinske analize podataka ima životni ciklus koji se sastoji od šest faza, kao što je prikazano na slici 1. Također treba uzeti u obzir da je slijed faza prilagodljiv, tj. sljedeća faza u slijedu često ovisi o rezultatima iz prethodne faze. Najznačajnije zavisnosti između faza označene su strelicama. Na primjer, pretpostavimo da smo u fazi modeliranja. Ovisno o ponašanju i karakteristikama modela, možda ćemo se morati vratiti u fazu pripreme podataka za daljnje poboljšavanje prije nego prijeđemo na fazu evaluacije modela. Pitanja koja se susreću tijekom faze evaluacije mogu vratiti analitičara u bilo koju od prethodnih faza kako bi se poboljšale. Iterativnu prirodu CRISP-DM-a simbolizira vanjska kružnica na slici 1.



Slika 1. Faze CRISP-DM

Faze CRISP-DM-a:

1. Faza razumijevanja poslovanja/istraživanja
    a. Jasno navesti ciljeve i zahtjeve projekta u smislu poslovne ili istraživačke jedinice
    b. Prevesti ove ciljeve i ograničenja u formulaciju definicije problema dubinske analize podataka
    c. Pripremiti preliminarnu strategiju za postizanje tih ciljeva
2. Faza razumijevanja podataka
    a. Prikupiti podatke
    b. Upotrijebiti istraživačku analizu podataka kako biste se upoznali s podacima i otkrili početne uvide
    c. Procijeniti kvalitetu podataka
    d. Ukoliko je potrebno, odabrati zanimljive podskupove koji sadrže uzorke podložne ispitivanju

3. Faza pripreme podataka
    a. Priprema seta podataka od početnih, sirovih, prljavih podataka do podataka koji će se koristiti za kasnije faze
    b. Odabir stanja i varijabli koje želimo analizirati i koji su prikladni za analizu
    c. Izvršiti transformacije nad određenim varijablama, ako je potrebno
    d. Očistiti neobrađene podatke tako da je set podataka spreman za alate za modeliranje
4. Faza modeliranja
    a. Odabrati i primijeniti odgovarajuće tehnike modeliranja
    b. Kalibrirati postavke modela za optimalne rezultate
    c. Moguće koristiti različite tehnike za isti problem dubinske analize podataka
    d. Vraćanje u fazu pripreme podataka kako bi se oblik podataka uskladio s posebnim zahtjevima određene tehnike dubinske analize podataka
5. Faza evaluacije
    a. Rezultat faze modeliranja je jedan ili više modela koji se moraju procijeniti za kvalitetu i učinkovitost prije nego se implementiraju
    b. Odrediti postiže li model ciljeve postavljene u fazi 1
    c. Ustvrditi je li neka važna strana poslovnog ili istraživačkog problema nedovoljno uračunata
    d. Donijeti odluku o upotrebi rezultata dubinske analize podataka
6. Faza implementacije
    a. Izrada modela ne predstavlja završetak problema, izrađene modele se treba iskoristiti
    b. Primjer jednostavne implementacije: izraditi izvješće
    c. Primjer složenije implementacije: provesti paralelni proces dubinske analize podataka u drugom odjelu tvrtke
    d. Za tvrtke, kupac često provodi implementaciju na temelju danog modela

Najčešći zadaci dubinske analize podataka su opis, procjena, predviđanje, razvrstavanje, grupiranje i udruživanje. Za naš daljnji rad zadatak je procjena koju ćemo dobiti različitim postupcima testiranja hipoteza.

2. Postupci testiranja hipoteze

Većina ljudi svakodnevno daju mnoge izjave. Također znanstvenici, inženjeri ili istraživači često imaju hipoteze o određenim dijelovima njihovog posla. Njihove hipoteze moraju biti potvrđene ili dokazane ili odbijene. Da bi se to ostvarilo prikupljaju se podaci pomoću kojih će potvrditi svoju hipotezu ili baciti određenu sumnju na nju. Ovaj proces se zove testiranje hipoteze. Testiranje hipoteze je primjena određenog skupa pravila koji nam pomažu u prihvaćanju ili odbijanju hipoteze. Metoda za prihvaćanje ili odbijanje hipoteze može se opisati u šest jednostavnih koraka:

1. Zauzeti početni stav – NULL hipoteza

NULL hipoteza najčešće zauzima stav da se ništa nije promijenilo. Simbol koji se koristi je H_0.

2. Odrediti alternativni stav – Alternativna hipoteza

Alternativna hipoteza, koja se zapisuje kao H_a, je suprotna početnom stavu, tj. zauzima stav da se nešto promijenilo.

3. Odrediti prikladan test

Ovisi o vrsti podataka na kojima vršimo testiranje.

4. Odrediti razinu značajnosti ili kritičnu p-vrijednost

Svako testiranje hipoteza je podložno greškama. Postoje dvije osnovne vrste grešaka:

1. Greška tipa 1 – Odbijanje H_0 iako je točna. Vjerojatnost ove greške se zapisuje simbolom α
2. Greška tipa 2 – Odbijanje H_a iako je točna. Vjerojatnost ove greške se zapisuje simbolom β

Zadatak pri testiranju svake hipoteze je odrediti razinu greške tipa 1. Kritičnu p-vrijednost određuje osoba koja odrađuje test te je to razina p-vrijednosti koja će mu reći je li uzorak dovoljno značajan u usporedbi s null hipotezom što ukazuje na to da bi se null hipoteza trebala odbaciti u korist alternativne hipoteze.

5. P-vrijednost ili kritično područje veličine α

P-vrijednost možemo naći u rasponu od 0.0 do 1.0. kako se približava 0.0 to pokazuje da je uzorak rijedak ishod u odnosu na populaciju kakvu smo postavili u hipotezi. Što je p-vrijednost bliža nuli to imamo veće dokaze protiv null hipoteze.

6. Izjava o zaključku

Odluka ovisi o p-vrijednost. Kad je p-vrijednost niska (manja od kritične p-vrijednosti), odbacujemo null hipotezu. Ukoliko je p-vrijednost visoka (veća od kritične p-vrijednosti), prihvaćamo null hipotezu. Zaključak bi također trebao biti, ukoliko je moguće, lišen statističkih pojmova i terminologije. U zaključku se kao statistički pokazatelj navodi razina značajnosti.

2.1. Test varijance ANOVA

ANOVA je parametarska metoda prikladna za usporedbu središnje vrijednosti za dvije ili više samostalnih uzoraka. ANOVA se izvodi izvršavanjem sljedeće usporedbe. Usporedite:

1. Varijabilnost između uzoraka, odnosno varijabilnost u uzorku središnjih vrijednosti
2. Varijabilnost unutar skupa uzoraka, tj. varijabilnost unutar svakog skupa uzoraka

Kada je (1) puno veći od (2), to predstavlja dokaz da uzorci nisu jednaki. Dakle, analiza ovisi o mjerenju varijabilnosti, dakle uvodi se pojam test varijance. Formule potrebne za izračun nalaze se u tablici 1.

Izvor varijacije	*Zbroj kvadrata*	*Stupnjevi slobode*	*Kvadrat središnje vrijednosti*	*F*
Obrada	$SSTR = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$	$df_1 = k - 1$	$MSTR = \frac{SSTR}{df_1}$	$F_{data} = \frac{MSTR}{MSE}$
Greška	$SSE = \sum (n_i - 1)s_i^2$	$df_2 = n_t - k$	$MSE = \frac{SSE}{df_2}$	
Ukupno	SST			

Tablica 1. Test varijance ANOVA

Ispitna statistika F_{data} bit će velika kada je varijabilnost između uzoraka mnogo veća od varijabilnosti unutar uzoraka, što ukazuje na situaciju koja zahtijeva odbacivanje null hipoteze. P-vrijednost je P ($F>F_{data}$); odbaciti null hipotezu kada je p-vrijednost mala, što se događa kada je F_{data} velika.

Kritična vrijednost se određuje razinom značajnosti (najčešće .05) i stupnjevima slobode (df_1 i df_2).

	df_1											
df_2	1	2	3	4	5	6	7	8	9	10	11	12
2	18.5	19.0	19.2	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.77	3.73	3.68	3.64	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.21	2.16
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.04
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83

Slika 2. Tablica kritičnih vrijednosti ANOVA test varijance

2.2. Hi-kvadrat test nezavisnosti

Hi-kvadrat test koristimo za određivanje značajnog odnosa između dvije nominalne (kategoričke) varijable. Učestalost svake kategorije za jednu nominalnu varijablu uspoređuje se s kategorijama druge nominalne varijable. Podaci se mogu prikazati u tablici gdje svaki red predstavlja kategoriju za jednu varijablu, a stupac predstavlja kategoriju za drugu varijablu.

	Varijabla B *Kategorija B.1*	*Varijabla B* *Kategorija B.2*	*Varijabla B* *Kategorija B.3*	
Varijabla A *Kategorija A.1*	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,1} + O_{1,2} + O_{1,3}$
Varijabla A *Kategorija A.2*	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,1} + O_{2,2} + O_{2,3}$
	$O_{1,1} + O_{2,1}$	$O_{1,2} + O_{2,2}$	$O_{1,3} + O_{2,3}$	Ukupno varijabli A = N

Tablica 2. Primjer prikaza vrijednosti dviju varijabli

Prvo moramo izračunati očekivane vrijednosti dviju nominalnih varijabli.

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,j} \sum_{k=1}^{r} O_{k,j}}{N}$$

Gdje je

- $E_{i,j}$ – očekivana vrijednost
- $\sum_{k=1}^{c} O_{i,j}$ – suma i-tog stupca
- $\sum_{k=1}^{r} O_{k,j}$ – suma k-tog retka
- N – ukupan broj

	Varijabla B Kategorija B.1	*Varijabla B Kategorija B.2*	*Varijabla B Kategorija B.3*
Varijabla A Kategorija A.1	$E_{1,1}$	$E_{1,2}$	$E_{1,3}$
Varijabla A Kategorija A.2	$E_{2,1}$	$E_{2,2}$	$E_{2,3}$

Tablica 3. Primjer prikaza očekivanih vrijednosti dviju varijabli

Nakon izračuna očekivane vrijednosti primjenjuje se sljedeća formula za izračunavanje vrijednosti Hi-kvadrat testa neovisnosti.

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Gdje je

- x^2 – hi-kvadrat test neovisnosti
- $O_{i,j}$ – promatrana vrijednost dviju nominalnih varijabli
- $E_{i,j}$ – očekivana vrijednost dviju nominalnih varijabli

Stupanj slobode računamo koristeći formulu $DF = (r-1)(c-1)$, gdje je

- DF – stupanj slobode
- r – broj redova
- c – broj stupaca

Kritična vrijednost se određuje razinom značajnosti (najčešće .05) i stupnjem slobode (DF).

	α					
v	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973

Slika 3. Tablica kritičnih vrijednosti Hi-kvadrat test

2.3. Pearsonov korelacijski koeficijent

Pearsonova korelacija je parametarska mjera linearne povezanosti između dvije numeričke varijable. Dvije varijable x i y linearno su korelirane ako je povećanje x povezano s povećanjem y ili sa smanjenjem y. Korelacijski koeficijent r kvantificira snagu i smjer linearnog odnosa između x i y. Prag za značaj korelacijskog koeficijenta r ovisi o veličini uzorka, ali i o samim podacima. Ako postoji veliki broj zapisa (preko 1000), čak i male vrijednosti r, kao što je -0.1 ≤ r ≤ 0.1 mogu biti statistički značajne.

Ukoliko imamo skup podataka {x1,...,xn} koji sadrži n vrijednosti i drugi skup {y1,...,yn} koji sadrži n vrijednosti, tada računamo r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Gdje je:

- n – veličina uzorka
- x_i, y_i - individualne vrijednosti indeksa i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – središnja vrijednost uzorka, analogno za $\bar{y}$

Stupanj slobode računamo koristeći formulu $DF = n - 2$, gdje je

- DF – stupanj slobode
- N – broj instanci

Kritična vrijednost se određuje razinom značajnosti (najčešće .05) i stupnjem slobode (DF).

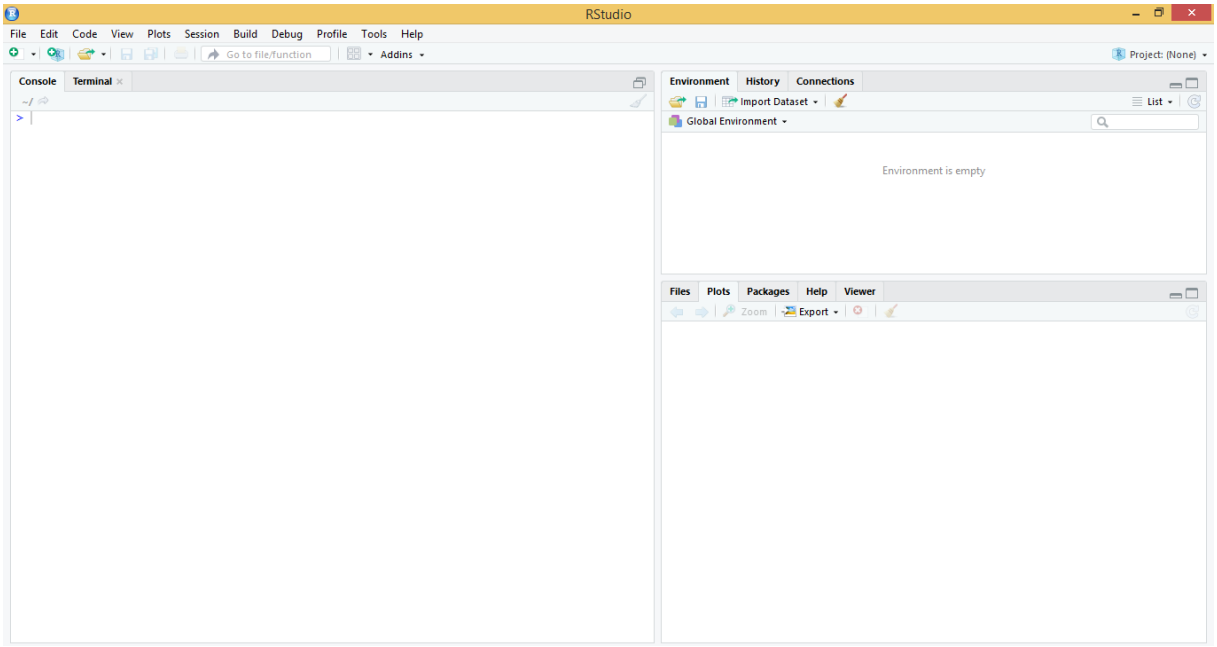
n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561

Slika 4. Tablica kritičnih vrijednosti za Pearsonov koeficijent

Tumačenje koeficijenta korelacije ovisi o kontekstu i svrsi. Ukoliko se verificira fizički zakon pomoću iznimno kvalitetnih instrumenata, korelacija od 0.8 može biti niska, no ista korelacija se može smatrati visoka u društvenim znanostima gdje imamo veći doprinos složenih čimbenika.

3. Projektni zadatak

Za ovaj dio rada izvest ću postupke statističkog testa na zadanom skupu podataka koristeći program Rstudio. Rstudio je integrirano razvojno okruženje za R. Uključuje konzolu, editor za isticanje sintakse koji podržava izravno izvršavanje koda, kao i alate za planiranje, povijest, ispravljanje pogrešaka i upravljanje radnim prostorom. Rstudio je dostupan u open source i komercijalnim izdanjima i radi na radnoj površini (Windows, Mac i Linux) ili u pregledniku povezanim s Rstudio Serverom ili Rstudio Server Pro (Debian/Ubuntu, RedHat/CentOS i SUSE Linux).



Slika 3. Sučelje na radnoj površini

3.1. Priprema podataka

Kako bih primijenila postupke testiranja hipoteze koji su mi zadani moram imati određeni skup podataka koji uključuju podatke na kojima mogu izvršiti sve zadane postupke testiranja. Skup podataka koji ću koristiti je Edu2mojprojekt koji sadrži 77 instanci i temelji se na podacima skupljenim u sustavu Mudri u okviru kolegija Programiranje (akademska godina 2013./2014.). Ovaj skup sam dobila pretprocesiran od strane profesorice i studenta koji je ovaj skup pripremio i također koristio za izradu svog diplomskog rada[3]. Skup Edu2mojprojekt sadrži sljedeće atribute:

Atribut	Opis
Id	Identifikator studenta (anonimizacija)
Lectures	Broj bodova koje je student dobio u okviru aktivnosti uz predavanja (max=7)
Quizzes	Ukupni bodovi koje je student dobio na oba kviza
Labs	Ukupni bodovi koje je student dobio na vježbama
Videos	Broj pokretanja video snimki sa snimljenim predavanjima
Self.Assesm.	Ukupan broj klikova u okviru samoprovjera – kao mjera aktivnosti
Grade	Diskretizirana ocjena studenta (FAIL, PASS, GOOD). FAIL uključuje F, PASS uključuje 2E i 2D, a GOOD uključuje A, B i C
Forum	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio Forumu uz demonstrature ili nije tijekom trajanja kolegija
Demos	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio demonstraturama ili nije tijekom trajanja kolegija
Red	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Red ili nije tijekom trajanja kolegija
Stog	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Stog ili nije tijekom trajanja kolegija
stabla1	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Uvod u stabla – ukratko ili nije tijekom trajanja kolegija
stabla2	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Uvod u stabla - drugi dio ili nije tijekom trajanja kolegija

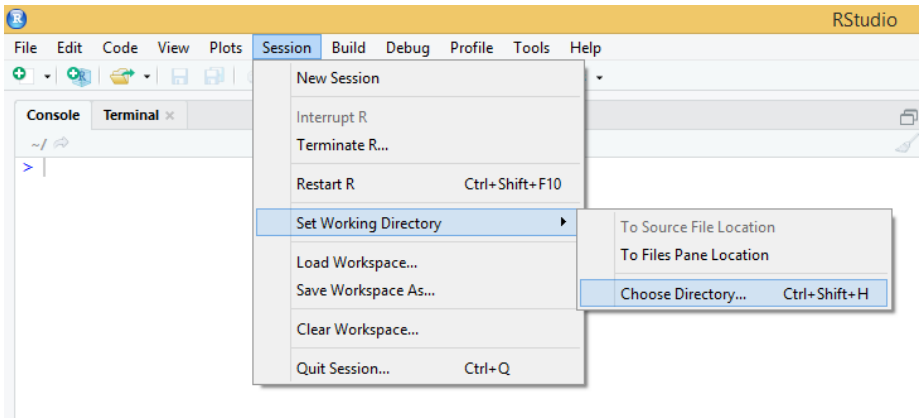
dinamičko	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Dinamičko programiranje ili nije tijekom trajanja kolegija
kružna	Sadrži logove 0 ili 1 ovisno o tome je li student pristupio prezentaciji Kružna lista ili nije tijekom trajanja kolegija

Tablica 4. Atributi iz skupa podataka Edu2mojprojekt

Kako bi ovaj skup imao sve potrebne atribute za sve postupke testiranja hipoteza koje ću implementirati, potrebno je dodatno ga urediti. Dodatno procesiranje podataka sam odlučila odraditi u Rstudiu.

3.2. Početak rada

Nakon otvaranja Rstudia treba se podesiti radni direktorij. To predstavlja direktorij iz kojeg će program povlačiti podatke tijekom rada.



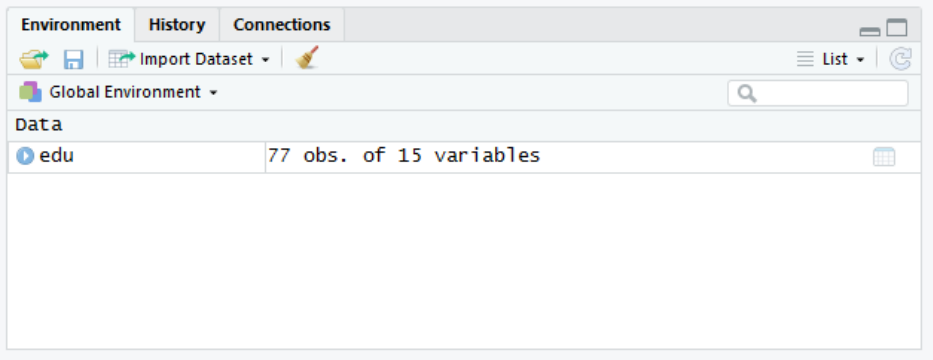
Slika 5. Odabir radnog direktorija

Sljedeći korak je uvoz skupa podataka u program za koji se koristi sljedeća naredba:

```
edu<-read.csv(file="edu2mojprojekt.csv", header=TRUE, sep=",")
```

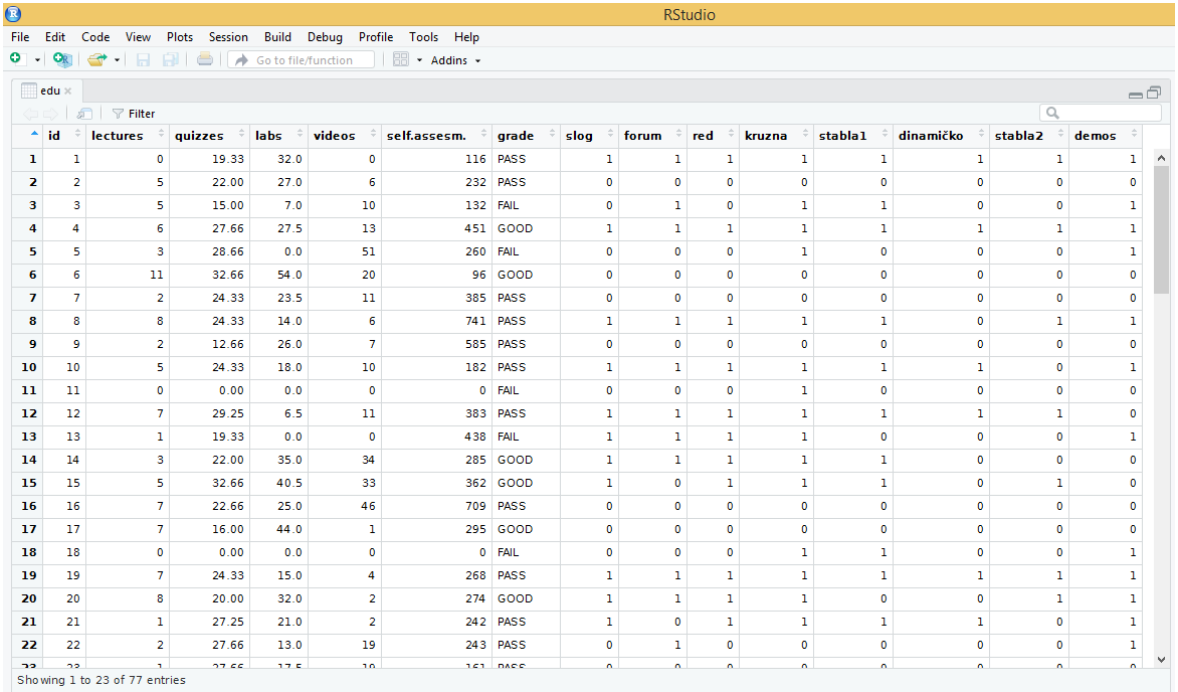
Naredba se sastoji od funkcije read.csv koja ima parametre naziv skupa podataka (u mom slučaju edu2mojprojekt.csv) i dodatnih parametara (u mom slučaju header="TRUE" i sep=","). Prvi dodatni parametar programu daje do znanja da je prvi redak u skupu podataka naziv varijabli, a drugi da su

podaci u skupu podataka odvojeni simbolom ",". Nakon uspješnog uvoza podataka u desnom dijelu sučelja programa Rstudio, naziv uvezenog skupa te informacije o broju instanci i broju varijabli pojavljuje se pod karticom „Environment".



Slika 6. Kartica "Environment"

Dvostrukim lijevim klikom na ime uvezenog skupa otvara se novi dijaloški okvir s detaljnim prikazom skupa podataka s nazivima instanci i varijabli.



	id	lectures	quizzes	labs	videos	self.assesm.	grade	slog	forum	red	kruzna	stabla1	dinamičko	stabla2	demos
1	1	0	19.33	32.0	0	116	PASS	1	1	1	1	1	1	1	1
2	2	5	22.00	27.0	6	232	PASS	0	0	0	0	0	0	0	0
3	3	5	15.00	7.0	10	132	FAIL	0	1	0	1	1	0	0	1
4	4	6	27.66	27.5	13	451	GOOD	1	1	1	1	1	1	1	1
5	5	3	28.66	0.0	51	260	FAIL	0	0	0	1	0	0	0	1
6	6	11	32.66	54.0	20	96	GOOD	0	0	0	0	0	0	0	0
7	7	2	24.33	23.5	11	385	PASS	0	0	0	0	0	0	0	0
8	8	8	24.33	14.0	6	741	PASS	1	1	1	1	1	0	1	1
9	9	2	12.66	26.0	7	585	PASS	0	0	0	0	0	0	0	0
10	10	5	24.33	18.0	10	182	PASS	1	1	1	1	1	1	0	1
11	11	0	0.00	0.0	0	0	FAIL	0	0	0	1	0	0	0	0
12	12	7	29.25	6.5	11	383	PASS	1	1	1	1	1	1	1	0
13	13	1	19.33	0.0	0	438	FAIL	1	1	1	1	0	0	0	1
14	14	3	22.00	35.0	34	285	GOOD	1	1	1	1	1	0	0	0
15	15	5	32.66	40.5	33	362	GOOD	1	0	1	1	1	0	1	0
16	16	7	22.66	25.0	46	709	PASS	0	0	0	0	0	0	0	0
17	17	7	16.00	44.0	1	295	GOOD	0	0	0	0	0	0	0	0
18	18	0	0.00	0.0	0	0	FAIL	0	0	0	1	1	0	0	1
19	19	7	24.33	15.0	4	268	PASS	1	1	1	1	1	1	1	1
20	20	8	20.00	32.0	2	274	GOOD	1	1	1	1	0	0	1	1
21	21	1	27.25	21.0	2	242	PASS	1	0	1	1	1	1	0	1
22	22	2	27.66	13.0	19	243	PASS	0	1	0	0	0	0	0	1

Slika 7. Detaljni prikaz skupa podataka edu

Za potrebe ovog diplomskog rada, prije testiranja hipoteza moram još dodatno urediti skup. Prvo moram dodati još kategorijskih varijabli, ali i još numeričkih varijabli te neke mogu izuzeti iz skupa.

Prvo izuzimam varijable koje neću koristiti.

```
edu<-edu[c(1:7)]
```

Nakon toga dodajem kategorijske varijable koje će prikazivati kategorije, broj bodova s predavanja, kviza i vježbi.

```
edu<-mutate(edu, lectures.factor=if_else(lectures<4, "LOW.lectures", "HIGH.
lectures"))
edu$lectures.factor=as.factor(edu$lectures.factor)


edu<-mutate(edu, quizzes.factor=if_else(quizzes<20, "LOW.quizzes", "HIGH.qu
izzes"))
edu$quizzes.factor=as.factor(edu$quizzes.factor)


edu<-mutate(edu, labs.factor=if_else(labs<30, "LOW.labs", "HIGH.labs"))
edu$labs.factor=as.factor(edu$labs.factor)
```

Nakon toga dodajem još jednu numeričku vrijednost, ukupni broj bodova.

```
edu<-mutate(edu, score=lectures + quizzes + labs)
```

Novo dodane vrijednosti u skup podataka:

Atribut	Opis
lectures.factor	Diskretizirani bodovi iz predavanja (HIGH.lectures, LOW.lectures). LOW.lectures uključuje studente koji su imali manje od 4 boda, ostali pripadaju kategoriji HIGH.lecture
quizzes.factor	Diskretizirani bodovi iz kvizova (HIGH.quizzes, LOW.quizzes). LOW.quizzes uključuje studente koji su imali manje od 20 boda, ostali pripadaju kategoriji HIGH.lecture
labs.factor	Diskretizirani bodovi iz vježbi (HIGH.labs, LOW.labs). LOW.labs uključuje studente koji su imali manje od 30 boda, ostali pripadaju kategoriji HIGH.labs
score	Ukupni broj skupljenih bodova (lectures + quizzes + labs)

Tablica 5. Novi atributi dodani u skup podataka edu

3.2.1. ANOVA

H_0:Broj pogledanih videa ne ovisi o bodovima iz (a. predavanja, b. kvizova, c. vježbi)

H_a:Broj pogledanih videa ovisi o bodovima iz (a. predavanja, b. kvizova, c. vježbi)

$\alpha = 0.05$

Broj pogledanih videa ne ovisi o bodovima iz predavanja

Unos	`videolectures.results<-aov(edu$videos ~ edu$lectures)`					
Unos	`summary(videolectures.results)`					
Ispis		Df	Sum Sq	Mean Sq	F value	Pr(>F)
	edu$lectures	1	651	650.6	6.487	0.0129
	Residuals	75	7522	100.3		
Zaključak	Ako je df_1=1 i df_2=75, a F=>3.96847 onda odbijamo null hipotezu. Broj pogledanih videa ovisi o bodovima s predavanja.					

Tablica 6. Test varijance ANOVA1

Broj pogledanih videa ne ovisi o bodovima iz kvizova

Unos	`videoquizzes.results<-aov(edu$videos ~ edu$quizzes)`					
Unos	`summary(videoquizzes.results)`					
Ispis		Df	Sum Sq	Mean Sq	F value	Pr(>F)
	edu$quizzes	1	1630	1629.8	18.68	3.69e-05
	Residuals	75	6543	87.2		
Zaključak	Ako je df_1=1 i df_2=75, a F=>3.96847 onda odbijamo null hipotezu. Broj pogledanih videa ovisi o bodovima iz kvizova.					

Tablica 7. Test varijance ANOVA2

Broj pogledanih videa ne ovisi o bodovima iz vježbi

Unos	`videolabs.results<-aov(edu$videos ~ edu$labs)`					
Unos	`summary(videolabs.results)`					
Ispis		Df	Sum Sq	Mean Sq	F value	Pr(>F)
	edu$labs	1	665	664.8	6.64	0.0119
	Residuals	75	7508	100.1		
Zaključak	Ako je df_1=1 i df_2=75, a F=>3.96847 onda odbijamo null hipotezu. Broj pogledanih videa ovisi o bodovima iz vježbi.					

Tablica 8. Test varijance ANOVA3

U sva tri slučaja je F-vrijednost veća od kritične vrijednosti te zaključujemo da broj pogledanih videa ovisi o bodovima iz predavanja, kvizova i vježbi.

3.2.2. Hi-kvadrat test nezavisnosti

H_0:Ocjena ne ovisi o bodovima iz (a. predavanja, b. kvizova, c. vježbi)

H_a:Ocjena ovisi o bodovima iz (a. predavanja, b. kvizova, c. vježbi)

$\alpha = 0.05$

Ocjena ne ovisi o bodovima s predavanja

Unos	`table(edu$grade,edu$lectures.factor)`		
Ispis		HIGH.lectures	LOW.lectures
	FAIL	3	22
	GOOD	20	2
	PASS	19	11
Unos	`chisq.test(table(edu$grade,edu$lectures.factor))$expected`		
Ispis		HIGH.lectures	LOW.lectures
	FAIL	13.63636	11.36364
	GOOD	12.00000	10.00000
	PASS	16.36364	13.63636
Unos	`chisq.test(table(edu$grade,edu$lectures.factor))`		

Ispis	`Pearson's Chi-squared test`
	
	`data:  table(edu$grade, edu$lectures.factor)`
	`X-squared = 30.92, df = 2, p-value = 1.931e-07`		

Zaključak	Ako je df=2, a x^2=>5.99147 onda odbijamo null hipotezu. Ocjena ovisi o bodovima s predavanja.		

Tablica 9. Hi-kvadrat test nezavisnosti 1

Ocjena ne ovisi o bodovima iz kvizova

Unos	`table(edu$grade,edu$quizzes.factor)`		
Ispis		HIGH.quizzes	LOW.quizzes
	FAIL	6	19
	GOOD	20	2
	PASS	23	7
Unos	`chisq.test(table(edu$grade,edu$quizzes.factor))$expected`		
Ispis		HIGH.quizzes	LOW.quizzes
	FAIL	15.90909	9.090909
	GOOD	14.00000	8.000000
	PASS	19.09091	10.909091

Unos	`chisq.test(table(edu$grade,edu$quizzes.factor))`

Ispis	`Pearson's Chi-squared test`
	
	`data: table(edu$grade, edu$quizzes.factor)`
	`X-squared = 26.245, df = 2, p-value = 1.999e-06`

Zaključak	Ako je df=2, a x^2=>5.99147 onda odbijamo null hipotezu. Ocjena ovisi o bodovima iz kvizova.

Tablica 10. Hi-kvadrat test nezavisnosti 2

Ocjena ne ovisi o bodovima iz vježbi

Unos	`table(edu$grade,edu$labs.factor)`		
Ispis		HIGH.labs	LOW.labs
	FAIL	0	25
	GOOD	17	5
	PASS	3	27
Unos	`chisq.test(table(edu$grade,edu$labs.factor))$expected`		
Ispis		HIGH.labs	LOW.labs
	FAIL	6.493506	18.50649
	GOOD	5.714286	16.28571
	PASS	7.792208	22.20779
Unos	`chisq.test(table(edu$grade,edu$labs.factor))`		

Ispis	`Pearson's Chi-squared test`
	
	`data: table(edu$grade, edu$labs.factor)`
	`X-squared = 42.863, df = 2, p-value = 4.924e-10`		

Zaključak	Ako je df=2, a x^2=>5.99147 onda odbijamo null hipotezu. Ocjena ovisi o bodovima iz vježbi.		

Tablica 11. Hi-kvadrat test nezavisnosti 3

U sva tri slučaja je x^2 veća od kritične vrijednosti te zaključujemo da ocjena studenta ovisi o bodovima iz predavanja, kvizova i vježbi.

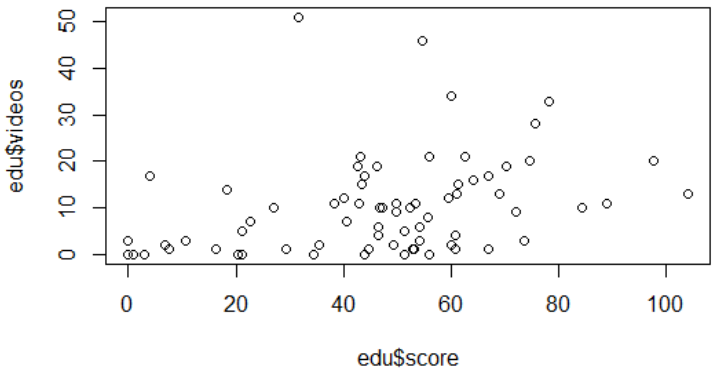
3.2.3. Pearsonov korelacijski koeficijent

H_0:Ukupni bodovi ne ovisi o broju pogledanih videa/samoprovjera

H_a: Ukupni bodovi ovisi o broju pogledanih videa/samoprovjera

$\alpha = 0.05$

Ukupni bodovi ne ovise o broju pogledanih videa

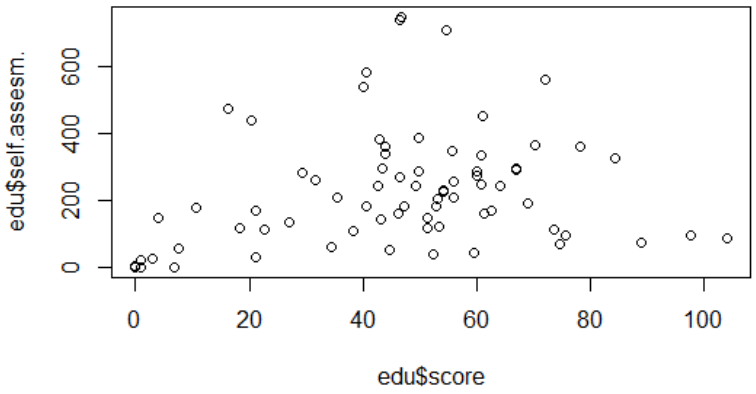
Unos	`plot(edu$score,edu$videos)`
Ispis	
Unos	`cor.test(edu$score,edu$videos)`

Ispis	`Pearson's product-moment correlation`
	
	`data:  edu$score and edu$videos`
	`t = 3.7178, df = 75, p-value = 0.0003851`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	` 0.1870290 0.5682531`
	`sample estimates:`
	`      cor`
	`0.3944828`

Zaključak	Ako je df=75, a r=>0.227 onda odbijamo null hipotezu. Pearsonov korelacijski koeficijent je r=0.3944828 pa možemo zaključiti da ukupni bodovi ovise o broju pogledanih videa.

Tablica 12. Pearsonov korelacijski koeficijent 1

Ukupni bodovi ne ovise o broju pogledanih samoprovjera

Unos	`plot(edu$score,edu$self.assesm.)`
Ispis	
Unos	`cor.test(edu$score,edu$self.assesm.)`

Ispis	`Pearson's product-moment correlation`
	
	`data:  edu$score and edu$self.assesm.`
	`t = 2.7765, df = 75, p-value = 0.006935`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	` 0.0872898 0.4954026`
	`sample estimates:`
	`      cor`
	`0.3052997`

Zaključak	Ako je df=75, a r=>0.227 onda odbijamo null hipotezu. Pearsonov korelacijski koeficijent je r=0.3052997 pa možemo zaključiti da ukupni bodovi ovise o broju odrađenih samoprovjera.

Tablica 13. Pearsonov korelacijski koeficijent 2

U oba slučaja je koeficijent r veći od kritične vrijednosti te zaključujemo da ukupni broj bodova ovisi o broju pregledanih videa i o broju odrađenih samoprovjera.

4. Zaključak

U ovom diplomskom radu provela sam postupke testiranja hipoteza koristeći se testom varijance ANOVA, Hi-kvadrat testom nezavisnosti i Pearsonovim korelacijskim koeficijentom. Iako sam za sva tri testiranja koristila isti skup podataka, nisam mogla uvijek testirati istu hipotezu zbog razlike u prirodi podataka koji za različita testiranja moraju biti numerički ili kategorički. Od sva tri testiranja smatram Pearsonov korelacijski koeficijent najintuitivnijim, pogotovo jer se i grafički može najlakše prikazati u RStudiu. Pearsonovim korelacijskim koeficijentom smo dokazali da ukupni broj bodova ovisi o broju pregleda videa ili samoprovjera iako su to jedini atributi u tablici koji ne utječu izravno na ukupi broj bodova pojedinog studenta. Hi-kvadrat testom smo provjeravali kategoričke podatke pa smo tako dokazali kako ocjena ovisi i o visini bodova na predavanjama i na kvizovima i na vježbama. To smo mogli očekivati s obzirom na to da se ocjena računa pomoću ova tri atributa. Također sam koristila test varijance ANOVA. Korištenjem testa varijance ANOVA sam zaključila da, iako broj pogledanih videa ne utječe izravno na bodove iz predavanja, kvizova i vježbi, da o njima ovisi.

Iako mi je provjeravanje hipoteza samo na temelju uzorka informacije u početku zvučalo sumnjivo, neke od mojih pretpostavki sam uspjela dokazati. Također smatram da bi za ovakvo testiranje na ovakvoj vrsti podataka sigurno bilo točnije kad bi se dodao veći broj studenata, tj. kad bi se spojili podaci o istom kolegiju, ali s više akademskih godina.

5. Literatura

1. Daniel T. Larose, Chantal D. Larose: Discovering knowledge in data (An intoduction to data mining), Hoboken, New Jersey, 2014.
2. Daniel T. Larose, Chantal D. Larose: Data mining and predictive analytisc, Hoboken, New Jersey, 2015.
3. Čanić Josip, Primjena logističke regresije u znanosti o podacima, Rijeka, 2017.
4. Memišević Haris, Biščević Inga: Statistički putokazi: Analiza varijanse (ANOVA) ili planirana komparacija - kako interpretirati podatke? Dohvaćeno 05.05.2016 na https://www.statistics.laerd.com
5. Rosalind L. P. Phang: Basic concepts in hypothesis testing, Dohvaćeno 11.09.2018 na https://sms.math.nus.edu.sg/smsmedley/Vol-16-2/Basic%20concepts%20in%20hypothesis%20testing(Rosalind%20L%20P%20Phang).pdf
6. Mock Thomas, A gentle gide to Tidy statistics in R (part 1 and part 2), Dohvaćeno 11.09.2018 na https://towardsdatascience.com/a-gentle-guide-to-statistics-in-r-a1da223e08b7

6. Tablice

7. Slike

8. Popis priloga

1. Programski kod napisan u R programskom jeziku za rješavanje projektnog zadatka u programu Rstudio.
2. Skup podataka Edu2mojprojekt.csv na kojem se izvršavalo testiranje hipoteza.