Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 656.8
DOI: https://doi.org/10.24867/06DS04Ninovic

PRIMENA SOFTVERA QGIS I PROGRAMSKOG JEZIKA PYTHON U ANALIZI PODATAKA

USE OF QGIS AND PYTHON FOR DATA ANALYSIS

Milena Ninović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – SAOBRAĆAJ

Kratak sadržaj – *Zadatak ovog istraživanja jeste da se za indikator Bruto domaćeg proizvoda (BDP) utvrdi vrednost korelacije sa drugim indikatorima i da se na taj način odredi koji od njih, i u kolikoj meri, utiču na porast ili smanjenje BDP-a. Za analizu korelacije BDP-a i sedam indikatora korišćen je program napisan u programskom jeziku „Python“. Osim toga, program je grupisao u CSV fajlove vrednosti za sve indikatore po godinama. Za indikator BDP-a , učitani su CSV fajlovi u programu QGIS, u kojem je prikazano kretanja BDP-a na mapi za sve zemlje Evrope tokom deset godina. Zaključeno je da indikatori „Ruralno“ i „Urbano stanovništvo“ imaju najznačajniji uticaj na BDP. Pokazano je da se pomoću ova dva programa može vršiti analiza velikog broja indikatora.*

Ključne reči:*Python, QGIS, Korelaciona analiza*

Abstract – *Python application was used for Gross Domestic Product (GDP) to indicator cooralation calculation. Python application also stored coorelation calculations to CSV files, having data values groped by year. GDP data was loaded in to QGIS from CSV files, which were than shown on Europe map for a span of ten year period. It was concluded that „Rurhal“ and „Urban Population“ have greatest impact on GDP. It was also shown how we can use software in order to analyse big data problems.*

Keywords: *Python, QGIS, correlation analysis*

1. UVOD

Ovo istraživanje bavi se veoma aktuelnom i kompleksnom problematikom analize i prezentacije podataka putem različitih programa. U ovom radu dat je primer kako se od velike količine podataka može izdvojiti samo izvestan broj koji se u ovom slučaju tiče vrednosti za osam indikatora, za deset godina i za zemlje Evrope. Potom se za željene indikatore može izračunati korelaciona analiza, takođe putem programa.

Nakon toga, dat je jednostavan prikaz kako se softver QGIS može iskoristiti za upotpunjivanje ove analize, ili prosto za vizuelizaciju kretanja nekih indikatora.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Dragana Šarac, vanr.prof.

2. PRIMENA PROGRAMSKOG JEZIKA PYTHON

Za potrebe ovog istraživanja prvobitno je bilo potrebno pribaviti neophodnu listu (bazu) podataka, a nakon toga, izvršiti njenu obradu programom napisanom u programskom jeziku „Python“. Podaci potrebni za analizu dobijeni su na sajtu: https://www.kaggle.com/, pod nazivom „World development indicator“.

„Python“ je interpretirani, objektno orijentisani jezik visokog nivoa, namenjen za pravljenje svih vrsta aplikacija – od inženjerskih i naučnih, do poslovnih i veb primena. Pravila jezika su jednostavna, a izvorni kod je često dosta kraći nego u slučaju Java ili C/C++ ekvivalenta [1].

Obrada je podrazumevala selekciju podataka za evropske države, ali i godine bitne za posmatranje i analizu.Podaci od interesa za ovo istraživanje tiču se vrednosti sedam indikatora o kojima će biti više reči u poglavlju 4. Za ovih sedam indikatora potrebno je izdvojiti vrednosti za evropske zemlje, od 2004. do 2013. godine. Ove godine su odabrane za analizu iz razloga što su u datoj listi podataka ponuđene vrednosti za brojne indikatore od 1960. do 2014. godine, s tim da za 2014. godinu nema podataka za veliki broj zemalja, te je pomenuti opseg od deset godina izabran kao najinteresantniji za analizu.

Koraci neophodni za manipulaciju potrebnim podacima biće prikazani u pseudo kodu. Koraci su sledeći:

Učitavanje biblioteka i podataka po državama

Da bi program uopšte mogao da radi, potrebno je učitati određene biblioteke. Ove biblioteke u sebi sadrže funkcije koje se pozivaju unutar koda. Učitavanje biblioteka se obavlja na sledeći način:

```
import naziv_biblioteke
import naziv_biblioteke as nb
```

Pozivanje samo jedne funkcije unutar neke biblioteke vrši se na sledeći način:

```
from naziv_biblioteke import funkcija1
```

Učitavanje podataka u CSV (*Comma Separated Values*) formatu izvršava se na sledeći način:

```
import csv
read_csv('./Naziv_fajla.csv')
```

CSV format je standardni format koji sadrži tabularne podatke u obliku teksta. Pogodan je za prenošenje podataka iz jedne aplikacije u drugu. Učitavanje CSV fajla je neophodno da bi se izgradio „Pandas DataFrame“ koji sadrži potrebne podatke.DataFrame predstavlja formatiranu matricu unutar koje se smeštaju podaci. DataFrame-ovi sadrže ključ u vidu ID rednog broja i nalaze se u bezimenoj koloni koja je na nultom mestu [2].

Učitavanje podataka za indikatore

Nakon prethodno izvršenih koraka, kao naredni, potrebno je izvršiti učitavanje podataka za indikatore za koje se vrši analiza. Učitavanje podataka obavlja funkcija pod nazivom „ucitavanjepod“. Primer poziva ove funkcije dat je u nastavku:

```
noviDataFrame =
ucitavanjepod('Naziv_indikatora')
```

Funkcija „ucitavanjepod“ pravi masku po filtru imena indikatora. Maska predstavlja listu koja unutar sebe sadrži vrednosti „True“ ili „False“ za dati uslov u „str.contains“ funkciji. Funkcija „str.contain“ proverava da li je svaki od podataka iz DataFrame-a jednak stringu „Naziv Indikatora“. Maska se definiše na sledeći način:

```
mask=data['Naziv_indikatora'].str.contains
(indicator)
```

Na osnovu maske, pravi se novi DataFrame pod nazivom „df“ koji sadrži samo tačne vrednosti iz maske:

```
df=data[mask]
```

Pored maske, potrebno je definisati i filtre koji će izdvajati vrednosti iz „df-a“, a gde će te vrednosti biti izdvojene za sledeće uslove:

- da se nalaze u listi evropskih država,
- da pripadaju određenim godinama.

Ovi uslovi, odnosno filtri, se definišu na sledeći način:

```
filtergodina = df['Year'] > 2003
filtergodina = df['Year'] < 2014
filtereu = df2['CountryCode'].isin(eu)
```

Na osnovu ovih komandi, podaci koji će se isfiltrirati važiće za vrednosti evropskih zemalja, za period od 2004. do 2013. godine.

Nakon ovoga, formira se novi DataFrame pod nazivom „df3“ unutar kojeg se nalaze isfiltrirane vrednosti. Ovo se vrši na sledeći način:

```
df3 = df2[filtergodina & filtereu]
```

Potom je potrebno napraviti novi DataFrame pod nazivom „dfnew“ koji unutar sebe poseduje tri kolone pod nazivom „CountryCode“, „Year“, „Value“ u koje će se smeštati neophodni podaci koji su prethodno isfiltrirani. Ova naredba se vrši sledećim ispisom:

```
Dfnew =
df3[['CountryCode','Year','Value']]
```

Pozivanjem funkcije „ucitavanjepod“, linija koda `return dfnew` vraća korisniku izmaskirani, isfiltrirani DataFrame, sa potrebnim kolonama. Za svaki indikator, ova funkcija se poziva posebno.

Skaliranje podataka

Za analizu ovog rada potrebno je posmatrati promene vrednosti nekog indikatora za svaku državu posebno, a kada bi se posmatrale sve države istovremeno, sa neskaliranim podacima, države sa visokim vrednostima nekog indikatora, u odnosu na one države koje takve vrednosti nemaju, uticale bi značajno više na ukupan rezultat, pa prezentacija takvih vrednosti ne bi bila u potpunosti realna.

Za indikatore za koje je skaliranje podataka potrebno, poziva se unapred definisana funkcija „scale“ na sledeći način:

```
indikator_scale = scale(indicator)
```

Funkcija „scale“ izdvaja podatke po indikatoru, po državama, i za svaku dobijenu listu podataka pronađe najveći član, a zatim sve elemente liste podeli najvećim članom. Novodobijene vrednosti vraća u prvobitni DataFrame.

Računanje korelacije i crtanje grafika

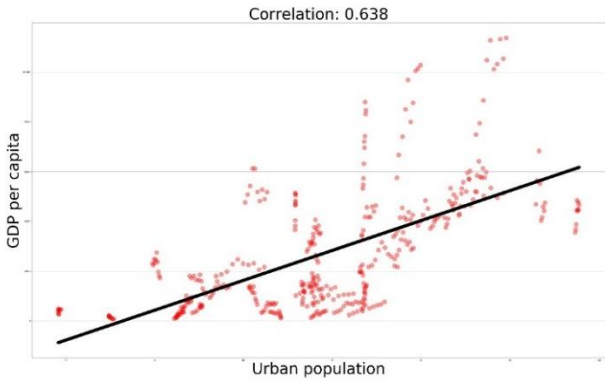
Računanje korelacije se vrši pomoću funkcije „corrcoef“ na sledeći način:

```
coef_korelacije =
np.corrcoef(indikator1,indicator2)
```

Crtanje grafika vrši se pomoću funkcije „axis.scatter“. Poziv funkcije dat je u nastavku:

```
axis.scatter(indicator1, indicator2,
color='Red', s = 450, alpha = 0.4)
```

Ova funkcija crta parove vrednosti za dva indikatora. Promenom parametara unutar zagrada se može uticati na izgled samog grafika. Primer grafika dat je na slici 1. Više reči o samom pojmu korelacije biće u poglavlju 4.



Slika 1. *Primer dobijenog grafika korelacije*

Linija regresije se računa pozivom funkcija „np.polyfit“ i „np.poly1d“. Ova linija se crta pomoću funkcije „plot“. Poziv funkcije dat je u nastavku:

```
zz = np.polyfit(list_urban,list_int,1)
pp = np.poly1d(zz)
```

3. PRIMENA SOFTVERA QGIS

Dovoljna je pomisao o tome koliko je samo vremena potrebno da se za neku lokaciju na karti pronađu i izdvoje svi podaci, da se svrstaju u razne tabele, da se potom povežu, usklade i, konačno, analiziraju. Zbog zadovoljavanja ovakvih potreba pogodno je formirati informacione sisteme koji će imati podatke o prostoru, koji će pratiti određena stanja prostora i koji će pomagati pri kontroli i upravljanju prostorom.

Da bi se formirao takav sistem, neophodno je da se obezbede kvalitetni kadrovi, računarska infrastruktura, podaci i odgovarajuće baze podataka. Sve ovo se može nazvati jednim imenom – Geografski informacioni sistem (GIS) [3].

Quantum GIS (QGIS) je besplatna aplikacija za GIS, za desktop platformu sa otvorenim kodom (eng. *open-source*) koja omogućava pregled, uređivanje i analizu geoprostornih podataka. QGIS funkcioniše kao softver za GIS, omogućavajući korisnicima da analiziraju i uređuju prostorne informacije, pored sastavljanja i eksporta grafičkih mapa [4].

Za prikaz kretanja vrednosti BDP-a tokom deset godina, za države Evrope, bilo je potrebno pribaviti određene karte sa kojima se može manipulisati u softveru QGIS. Kako se na sajtu „Poljoprivrednog fakulteta“ u Novom Sadu, na „Odseku za Geografske informacione sisteme“ nude određene, već gotove mape, jedna od tih mapa sveta iskorišćena je u svrhe ovog istraživanja. S obzirom na to da je predmet ovog istraživanja prikaz vrednosti BDP-a za države Evrope, potrebno je izvršiti uklanjanje svih ostalih zemalja iz atributne tabele. Tako „prečišćena“ mapa pogodna je za dalju manipulaciju i analizu. Njoj će biti pridružen CSV fajl koji će sadržati podatke u vidu vrednosti BDP-a.

Da bi se postiglo prikazivanje vrednosti BDP-a tokom deset godina, mora postojati deset različitih mapa sa mogućnošću da se one prikazuju pojedinačno. Svakoj mapi biće pridružen CSV fajl koji sadrži vrednosti za određenu godinu. Način na koji se dobijaju pomenute vrednosti, dat je u poglavlju 2. Da bi CSV fajl mogao da bude priključen QGIS-u, i da bi se ostvario odgovarajući prikaz mape koja je pogodna za dalju analizu, potrebno je izvršiti sledeće funkcije:

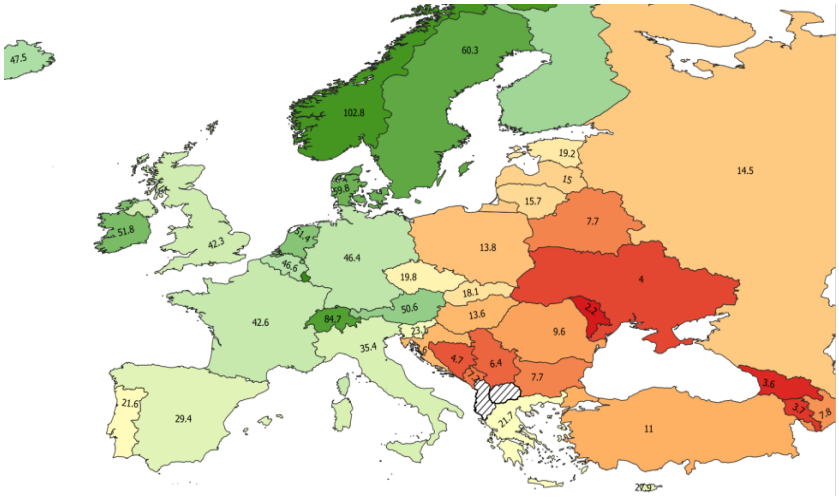
1. Dodati novi sloj sa tekstualnim podacima pomoću opcije „Add Delimited text“;
2. Nov sloj pripojiti karti na kojoj će se prikazivati vrednosti BDP-a. Podaci iz CSV fajlova (koji se „ubacuju“ u softver QGIS) bivaju pridruženi već postojećoj atributnoj tabeli koja sadrži indekse i nazive zemalja;
3. Izvršiti podešavanja prikaza vrednosti BDP-a, kao i podešavanja obojenosti država na osnovu ovih vrednosti.

Odabirom funkcije za prikaz statistike, koja je zapravo vrednost BDP-a po zemljama, dobija se mogućnost rangiranja vrednosti ovog indikatora. U skladu sa tim, vrednosti se rangiraju od najmanje ka najvećoj, gde su zemlje sa najmanjim vrednostima obojene najtamnijom nijansom crvene boje, a zemlje sa najvećim vrednostima su prekrivene najtamnijom nijansom zelene boje.

Dakle, što je nijansa crvene tamnija, to je vrednost BDP-a manja, a što je vrednost BDP-a veća, to je nijansa zelene tamnija. Zemlje za koje ne postoje podaci u atributnoj tabeli su išrafirane, a to su u ovom slučaju Albanija i Makedonija.

Zahvaljujući prethodno izvršenim podešavanjima, kao rezultat, moguće je pratiti kretanje BDP-a po evropskim zemljama, pojedinačno, od 2004. do 2013. godine (na slici2 je dat primer za 2013. godinu).

Tako se postiže uporedni prikaz i analiza podataka od interesa. Ovim pristupom može se omogućiti prikaz i drugih indikatora, ne samo BDP-a, ukoliko bi tako nešto bilo potrebno.



Slika 2. *Prikaz vrednosti BDP-a za 2013. godinu za zemlje Evrope u QGIS-u*

4. ANALIZA KORELACIJE

Korelaciona analiza ima za cilj da prikaže povezanost između promenljivih vrednosti, odnosno, da li postoji zavisnost među odabranim indikatorima. Vrednost korelacije utvrđuje se merenjem koeficijenta korelacije koji predstavlja numeričku vrednost kojom se označava stepen povezanosti između dve promenljive pojave. Ova vrednost se kreće od -1 do +1[5].

Identifikacija i analiza indikatora predstavlja važnu fazu u ovom radu.Cilj je da se utvrdi da li postoji visoka korelacija između pojedinih indikatora i BDP-a, odnosno da se utvrdi koji indikatori i u kolikoj meri utiču na porast ili smanjenje BDP-a. Za indikatore koji nisu srodni, ili nemaju logički međusobni uticaj, nije utvrđena korelacija. Predmet istraživanja su vrednosti indikatora merene u periodu od 2004. do 2013. godine. Rezulatati analize korelacije su predstavljeni kombinacijom grafičkog prikaza i numeričke vrednosti i obračunati su pomoću fukcija napisanih u programskom jeziku Python. Prikaz jednog od sedam grafika dat je na slici broj 1.

U skladu sa prethodno navedenim, može se izvršiti tumačenje dobijenih korelacija između različitih

indikatora. U tabeli 1. prikazani su parovi indikatora sa vrednostima njihove korelacije.

Tabela 1. *Skala korelacije za parove indikatora*

Indikator Prvi	Indikator Drugi	Vrednost Korelacije	Skala Korelacije
GDP per capita	Population, total	-0,232	zanemarljiva
GDP per capita	Urban population	0,638	srednja pozitivna
GDP per capita	Rural population	-0,638	srednja negativna
GDP per capita	Trade	0,327	slaba pozitivna
GDP per capita	Industry	-0,241	zanemarljiva
GDP per capita	Services	0,475	slaba pozitivna
GDP per capita	Labor force participation rate for ages 15-24	0,57	slaba pozitivna

5. ZAKLJUČAK

Analizom indikatora BDP-a, koji je jedan od pokazatelja životnog standarda, može se zaključiti kakva je razvijenost jedne zemlje i regiona. Iz tog razloga je baš ovaj indikator uzet kao primer za analizu u ovom istraživanju. Što se tiče Srbije, u posmatranom periodu od 2004. do 2013. godine, najveći BDP je imala tokom 2008. godine, što se može videti u dobijenim mapama u QGIS-u

Korelaciona analiza različitih indikatora pruža mogućnost da se utvrdi njihova veza sa rastom ili padom BDP-a. Radom programa napisanog u programskom jeziku Python utvrđeno je da indikatori „Ruralno“ i „Urbano stanovništvo“ imaju najznačajniji uticaj na BDP. Uopšteno, ukoliko postoji značajna korelacija nekih indikatora, može se izvršiti dalja analiza, tako što će se izmeriti korelacija sa nekim drugim indikatorima.

Uticajem i investiranjem u određeni indikator, indirektno se poboljšavaju vrednosti drugih indikatora koji su visoko i pozitivno korelisani sa tim indikatorom. Ograničenje ovog rada je to što su uzeti samo neki od indikatora, kao primer kako je moguće izvršiti ovakvu analizu. Rezultati ne moraju biti pogrešni zbog ovog ograničenja, već ne prikazuju međuzavisnost svih indikatora.

Rezultat ovog istraživanja pokazao je da se velike liste podataka relativno brzo i jednostavno mogu prilagoditi za korišćenje. Istovremeno, pokazana je prednost upotrebe programskih jezika u analizi podaka i njihovoj prezentaciji.

Programom QGIS postiže se bolje vizuelno sagledavanje statistike kretanja BDP-a, po godinama, sloj po sloj. Iako je u ovom radu, upravo na ovaj način, analiziran uticaj sedam indikatora na porast ili smanjenje BDP-a, ovakav pristup se može iskoristiti i u mnogim drugim slučajevima, kako za „čiščenje“ podataka, analizu korelacije, tako i za prikazivanje raznih indikatora i podataka putem karte.

6. LITERATURA

[1] M. Kovačević, „Osnove programiranja u Pajtonu“, Univerzitet u Beogradu – Građevinski fakultet, Beograd, Akademska misao, Beograd, 2017.

[2] Z. Hercigonja„Vizualizacija postupaka rješavanja programskih zadataka u programskom jeziku Python“, Imbriovec Jalžabetski 45c, 2017.

[3] Z. Čekerevac, S. Anđelić, S. Glumac, N. Dragović, „Savremene tendencije primene GIS tehnologija“

[4] QGIS (dostupno na: https://qgis.org/en/site/, Pristupano: 18. 08. 2019).

[5] D.Vuković, „Korelaciona analiza indikatora regionalne konkurentnosti: Primer Republike Srbije“, Ekonomski horizont, Godište XV, Univerzitet u Kragujevcu, Ekonomski fakultet, Kragujevac, Sveska 3, pp. 197-211, 2013.

Kratka biografija:

Milena Ninović rođena je u Valjevu 1995. god. Osnovne akademske studije upisala je 2014. godine i završila 2018. godine na smeru Poštanski saobraćaj i telekomunikacije. Master rad na Fakultetu tehničkih nauka iz oblasti Geografski informacioni sistemi odbranila je 2019.god.
kontakt: ninovic95@gmail.com