Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/17BE08Damjanovic

GRAMATIKA GRAFIKE ZA VIZUALIZACIJU PODATAKA I NJENE IMPLEMENTACIJE

GRAMMAR OF GRAPHICS FOR DATA VISUALIZATION AND ITS IMPLEMENTATIONS

Milica Damjanović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Grafički prikaz podataka ima značajnu ulogu u njihovoj analizi. U ovom radu su opisani osnovni koncepti gramatike grafike koja se uspešno koristi za vizualizaciju podataka. Predstavljene su komponente slojevite gramatike, slojevi i odgovarajuća preslikavanja podataka. Na primeru, implementiranom u programskom jeziku Python, demonstrirana je efektivna realizacija vizualizacije nad višedimenzionalnim podacima..*

Ključne reči: *gramatika grafike, vizualizacija, Python, analiza podataka*

Abstract – *Graphical data representation has significant role in its analysis. In this theses, basic concepts of grammar of graphics for data visualization are described. Research outline and explain components of layered grammar, layers and related data mappings. Example, implemented in Python programming language, demonstrate developing of efficient visualizations on multidimensional data.*

Keywords: *Grammar of graphics, visualization, Python, data analysis*

1. UVOD

Vizualizacija podataka predstavlja najoptimalniji način sortiranja i predstavljanja kompleksnih podataka. Jednostavan grafički prikaz može da bude koristan i da uštedi sate istraživanja. Grafici su laki za čitanje i interpretaciju, a ako su iz pouzdanih izvora, takođe ih možemo smatrat tačnim. Glavni cilj vizuelizacije podataka nije da čini podatke lepšim, nego pružanje korisnicima dolazak do skrivenih zaključaka u podacima, predstavljajući im ključne aspekte na intuitivniji, smisleniji i koncizniji način.

Vizualizacija podataka nije novina. Istorija vizualizacije potiče još pre 2.500 godina, kad nisu postojali ni računari ni alati za analizu sa vizuelnim rešenjima. Jedna od prvih poznatih vizualizacija je vizuelizacija koja prikazuje izbijanje kolere u ulici Broad Street u Londonu 1854. godine. Istraživač je pokušao da objasni statističkim podacima, a zatim i mapom, povezanost između kvaliteta izvora vode i slučajeva kolere. Mapa je pokazala da se najviše slučajeva kolere dešava u blizini pumpi sa vodom. Smatralo se da se kolera širi vazduhom, ali svojim vizuelnim prikazom podataka je uspeo da prikaže gde zaista nastaje problem.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Dunja Vrbaški, docent

2. RAZUMEVANJE GRAMATIKE GRAFIKE

U osnovi, gramatika grafike je okvir koji sledi slojevit pristup za opisivanje i konstruisanje vizualizacija ili grafike na strukturiran način. Vizualizacija koja uključuje višedimenzionalne podatke često ima više komponenti ili aspekata, a korišćenje slojevite gramatike grafike pomaže da se opiše i razume svaka komponenta koja je uključena u vizualizaciju u smislu podataka, estetike, razmere i objekata.

2.1. Značajne ličnosti koje su doprinele razvoju gramatike grafike

Originalnu gramatiku grafičkog okvira predložio je Leland Vilkinson [1] koja detaljno pokriva sve glavne aspekte koji se odnose na efikasnu vizualizaciju podataka. Vilkinson je početkom 1980-ih napisao SYSTAT, statistički programski paket. Ovaj je program poznat po svojoj sveobuhvatnoj grafici, uključujući prvu softversku implementaciju zaslona mapa topline (eng. heatmap), koji se danas široko koristi među biolozima. Vilkinsonova gramatika daje sažeta i sveobuhvatna pravila opisivanja objekata statističkih grafika koji uključuju podatke, transformaciju, skalu, koordinate, elemente, vodiče i prikaze kao i kako iscrtati ove objekte pomoću sistema iscrtavanja (slika 2). Njegova knjiga "Gramatika grafike" je takođe poslužila kao temelj za paket R ggplot2. Takođe se koristiti varijanta ovoga okvira, poznata kao slojevita gramatika grafičkog okvira, koju je predložio Hedli Vikam, renomirani istraživač iz oblasti nauke o podacima i osnivač poznatog R paketa za vizualizaciju ggplot2. Gramatika se razlikuje od Vilkinsonove po rasporedu komponenti, razvoju hijerarhije zadatih podešavanja, te zato što je ugrađen u drugi programski jezik.

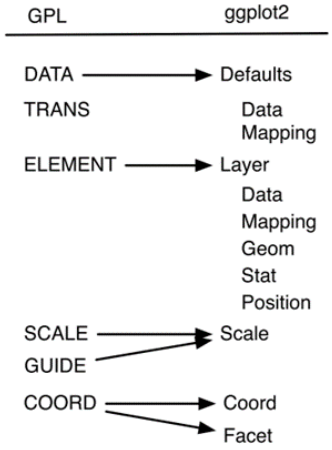
2.2. Komponente slojevite gramatike

Vikamova slojevita gramatika grafike koristi nekoliko slojevitih komponenti za opis bilo koje grafike ili vizualizacije. Najvažnije je da ima neke varijacije u odnosu na originalnu gramatiku grafike koju je predložio Vilkinson. Preciznije, slojevita gramatika definiše komponente grafika kao skup sledećih elemenata:

- Podrazumevani skup podataka i skup preslikavanja promenljivih u estetiku.
- Jedan ili više slojeva, svaki sastavljen od geometrijskog objekta, statističkih transformacija, i podešavanje položaja, i opciono, skup podataka i preslikavanja estetike.
- Jedna skala za svako korišćeno preslikavanje estetike.
- Koordinatni sistem.
- Specifikacija podele pogleda.

Ove komponente visokog nivoa su prilično slične komponentama iz Vilkinsonove gramatike.

U obe gramatike komponente su nezavisne, što znači da možemo generalno menjati jednu komponentu izolovano. Postoji više razlika unutar pojedinačne komponente, koje će biti opisane u detaljima koji slede.



Slika 1. *Mapiranje između komponenti Vilkinsonove gramatike (levo) i slojevite gramatike (desno)*

2.3. Slojevi

Slojevi su odgovorni za stvaranje objekata koje opažamo na grafiku. Sloj je sastavljen od četiri dela: mapiranje podataka i estetike, statistička transformacija (stat), geometrijski objekat (geom) i podešavanje položaja.

Obično svi slojevi na grafiku imaju nešto zajedničko, što je tipično jer su oni različiti pogledi na iste podatke. Sloj je ekvivalent Vilkinsonovom elementu. Međutim, parametrizacija je prilično drugačija. U Vilkinsonovoj gramatici svi delovi elementa su isprepleteni, dok su u slojevitoj gramatici oni odvojeni

2.4. Podaci i mapiranje

Podaci su očigledno kritičan deo grafike, ali važno je napomenuti da su nezavisni od ostalih komponenti: može se konstruisati grafik koja se može primeniti na više skupova podataka. Podaci su ono što apstraktnu grafiku pretvara u konkretnu.

Uz podatke, potrebna je specifikacija koje se promenljive preslikavaju na koju estetiku. Izbor dobrog preslikavanja je ključan za generisanje korisne grafike.

2.5. Statistička transformacija

Statistička transformacija (*stat*) transformiše podatke, obično ih sažimajući na neki način. *Stat* uzima skup podataka kao ulaz i vraća skup podataka kao izlaz, pa *stat* može dodati nove promenljive u originalni skup podataka. Moguće je preslikati estetiku na ove nove promenlljive. Geometrijski objekti, ili skraćeno geomi, kontrolišu vrstu grafika koji se kreira. Na primer, korišćenje tačkastog geoma će stvoriti raspršeni grafik, dok će upotreba linijskog geoma stvoriti linijski grafik. Geome možemo klasifikovati prema njihovoj dimenzionalnosti:

- 0d: tačka, tekst
- 1d: putanja, linija (uređena putanja)
- 2d: poligon, interval

Ponekad se mora prilagoditi položaj geometrijskih elemenata na grafiku kako u suprotnom ne bi prekrili jedan drugog. Vilkinson naziva ovo modifikatorima sudara.

2.6. Skale

Skala kontroliše preslikavanje podataka na atribute estetike, pa je potrebna jedna skala za svako svojstvo estetike koje se koristi u sloju. Skale su zajedničke po slojevima kako bi se osiguralo konzistentno preslikavanje podataka na estetiku.

Skala slojevite gramatike ekvivalentna je *skali (scale)* i *vodiču* (*guide*) Vilkinsonove gramatike. Postoje dve vrste vodiča: vodiči za skale i vodiči za beleške. U slojevitoj gramatici, vodiči se u velikoj meri automatski crtaju na osnovu opcije isporučene na relativnoj skali. Vodiči za beleške nisu neophodni ukoliko mogu biti kreirani sa kreativnom upotrebom geometrijskih objekata ako zavise od podataka, ili ako se osnovim sistemima za crtanje može se direktno pristupiti.

Skale se takođe računaju drugačije ukoliko je moguće preslikavanje promenljive proizvedene statistikom u estetiku. Ovo zahteva dva prolaza skaliranja, pre i posle statističke transformacije.

2.7. Koordinatni sistemi

Koordinatni sistem preslikava položaj objekata na ravan grafika. Položaj je često određen sa dve koordinate (x, y), ali može biti određen bilo kojim koordinatama. Dekartov koordinatni sistem je najčešći koordinatni sistem za dve dimenzije, dok se polarne koordinate i razne projekcije mape ređe koriste. Za veće dimenzije postoje paralelne koordinate (projektivna geometrija), mozaičke grafike (hijerarhijski koordinatni sistem) i linearne projekcije na ravan. Koordinatni sistemi utiču na sve promenljive položaja istovremeno i razlikuju se od skala u tome što menjaju i izgled geometrijskih objekata.

2.8. Podela pogleda

Postoji i još jedna značajna komponenta koja se pokazala dovoljno korisnom i koju bi trebalo uključiti u opštem okviru: podela pogleda (aspektovanje, eng. facet). Ovo olakšava stvaranje malih višekratnika različitih podskupova čitavog skupa podataka. Ovo je moćan alat pri istraživanju da li obrasci važe za sve uslove. Specifikacija podele pogleda opisuje koje promenljive treba koristiti za podelu podataka i kako ih treba rasporediti u mrežu. U Vilkinsonovoj gramatici, podela pogleda predstavlja aspekt koordinatnog sistema, sa donekle komplikovanom parametrizacijom: promenljiva podele navedena je unutar komponente *element* i zasebni *coord* parametar specificira da koordinatni sistem treba da bude podeljen ovom promenljivom. Ovo je pojednostavljeno u slojevitoj gramatici jer se podelapogleda vrši nezavisno od sloja i unutrašnjeg koordinatnog sistema. Manje je fleksibilno, jer se izgled pogleda uvek javlja u Dekartovom koordinatnom sistemu, ali u praksi nije ograničavajući.

2.9. Ugrađena gramatika

Prednosti ugrađivanja gramatike grafike u drugi programski jezik su očigledne: čovek odmah dobija sve postojeće sposobnosti tog jezika. Mogu se koristiti sve mogućnosti

potpunog programskog jezika za automatizaciju ponavljajućih zadataka: petlje za iteraciju preko promenljivih ili podskupova, promenljive za skladištenje često korišćenih komponenti i funkcija za enkapsuliranje i dekompoziciju uobičajenih zadataka. Nedostaci ugrađivanja gramatike su nešto suptilniji i u vezi su gramatičkim ograničenjima koja primenjuje jezik domaćin. Jedna od najvažnijih karakteristika gramatike je njena deklarativna priroda. Za očuvanje ove prirode u paketu ggplot2 koristi se operator + za kreiranje grafika dodavanjem delova zajedno. Funkcija ggplot stvara osnovni objekat, kome se dodaje sve ostalo. Ovaj osnovni objekat nije neophodan u samostalnoj gramatici.

2.10. Implikacije slojevite gramatike

Tri zanimljiva aspekta primene gramatike:

- Histogram, koji preslikava visinu linije na promenljivu koja nije u originalnom skupu podataka, i postavlja pitanja parametrizacije i zadatih vrednosti.
- Polarne koordinate, koje generišu tortne grafike iz trakastih grafika.
- Promenljive transformacije i tri mesta na kojima se mogu dogodit vrednosti

Jedan od najkorisnijih nacrta za posmatranje jednodimenzionalnih (1-D) distribucija je histogram. Histogram je prilično poseban jer preslikava estetiku (visinu linije) u promenljivu stvorenu statistikom (broj binova) i otvara neka pitanja u vezi sa parametrizacijom i izborom podrazumevanih vrednosti. Jedan koordinatni sistem koji se vrlo često koristi u poslovnoj grafici je polarni koordinatni sistem, koristi se za izradu okruglih dijagrama i radarskih prikaza. Polarni koordinatni sistem parametriše dvodimenzionalna ravan u smislu ugla θ i udaljenosti od kooridnatnog početka, ili poluprečnika, r. Postoje tri načina za transformaciju vrednosti korišćenjem ggplot2 biblioteke: transformacijom podataka, transformacijom skala i transformacijom koordinatnog sistema. Transformisanje podataka ili skala proizvodi grafike koji izgledaju vrlo slično, ali ose (i linije mreže) su različite: sve ostalo ostaje isto. To je zato što statistika radi na podacima koji su bili transformisana skalom.

3. IMPLEMENTACIJA

Na primeru jednog skupa podataka su prikazane vizualizacije dobijene korišćenjem gramtike grafike za opis prikaza. Sve vizualizacije su rađene u programskom jeziku Python. Plotnine je implementacija gramatike grafike, zasnovana na ggplot2. Gramatika omogućava korisnicima da sastavljaju grafike eksplicitnim preslikavanjem podataka na vizuelne objekte koji čine grafik. Uvek se počinje učitavanjem i posmatranjem skupa podataka koji se želi analizirati i vizualizovati. Korišćen je skup podataka koji sadrži informacije o prodaji računara.

Atributi koji čine skup podataka su:

- Company – tip string, kompanija koja je proizvela računar
- Product – tip string, model i oznaka računara
- TypeName – tip string, kog tipa je računar
- Inches – numeričkog tipa, veličina ekrana
- ScreenResolution – tip string, rezolucija ekrana
- Cpu – tip string, centralna procesorska jedinica
- Ram – tip string, radna memorija
- Memory – tip sring, Hard Disk / SSD memorija
- Gpu – niz, grafička procesorska jedinica
- OpSys – tip string, operativni sistem
- Price_euros – numerički tip, cena izražena u eurima

3.1. 2-D vizualizacija podataka

PODACI: Izvor informacija

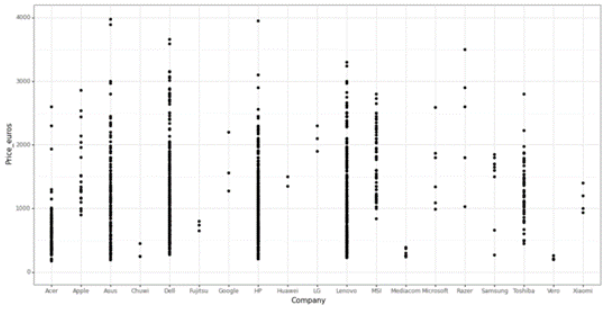
Prvi korak pri kreiranju vizualizacije podataka je određivanje podataka koji će se iscrtavati. Prilikom primene plotnine biblioteke kreira se ggplot objekat i prenosi se skup podataka koji se želi koristiti u konstruktor.

ESTETIKA: Definicija promeljive za svaku osu

Nakon što se navedu podaci koje je potrebno vizualizovati neophodno je definisati promenljive koja se žele koristiti za svaku osu na grafiku. Svaki red u okviru podataka može sadržati mnogo polja, pa se mora navesti koje promenljive se biraju za prikaz. Koristeći objekat ggplot, atribut kompanije se preslikava na horizontalnu grafičku osu, a cena izražena u eurima na vertikalnu osu (slika 2).

GEOMETRIJSKI OBJEKTI: Odabir različite vrste grafika

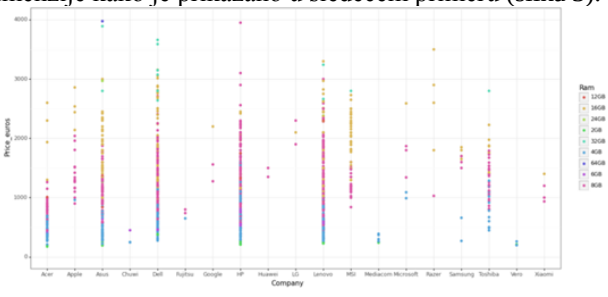
Nakon što se definišu svi podaci i atributi koji se žele koristiti na slici, mora se navesti geometrijski objekat koji će definisati način prikaza, odnosno kako treba iscrtati tačke podataka. Plotnine pruža mnogo geometrijskih objekata koji se mogu koristiti van okvira, poput linija, tačaka, šipki, poligona i još mnogo toga.



Slika 2. *Prikaz cene računara po kompanijama*

3.2. 3-D vizualizacija podataka

Za vizualizaciju tri dimenzije iz skupa podataka, možemo koristiti boju kao jednu od komponenti estetike za vizualizaciju jedne dodatne dimenzije pored druge dve dimenzije kako je prikazano u sledećem primeru (slika 3).

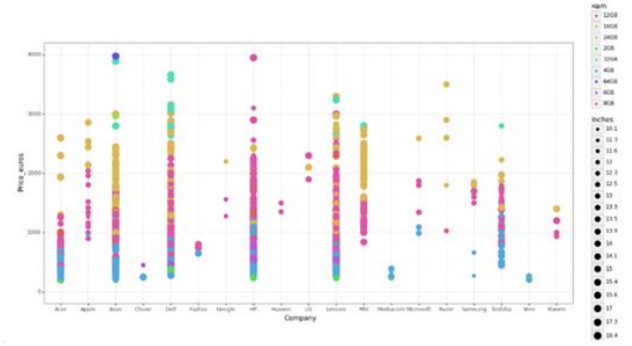


Slika 3. *Prikaz cene računara po kompanijama uz dodatnu dimenziju predstavljenu bojom*

3.3. 4-D vizualizacija podataka

Vizualizacija pokazuje koliko moćna estetika može biti u tome što pomaže da vizualizujemo više dimenzija

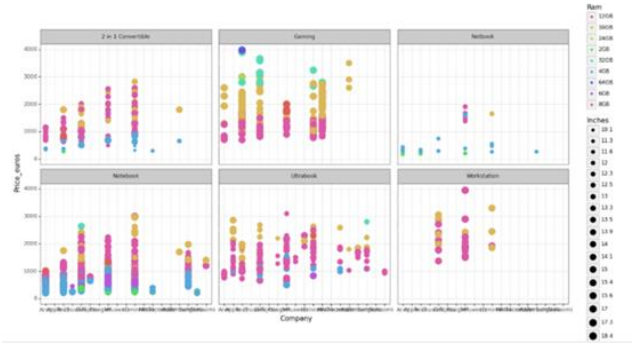
podataka na jednom grafiku. Da bismo vizualizovali četiri dimenzije iz skupa podataka, možemo upotrebiti boju i veličinu, kao dve estetike vizualizacije pored drugih regularnih komponenti. Za veličinu je izabran atribut koji nam daje informacije o veličini ekrana svakog računara u inčima (slika 4). Alternativno, može se takođe koristiti boja i podlea pogleda umesto veličine što je objašnjeno u predstavljanju primera 5-D vizualizacije.



Slika 4. *Prikaz cene računara po kompanijama uz dodatne dve dimenziju predstavljene bojom i veličinom*

3.4. 5-D vizualizacija podataka

Da bi se vizualizovali podaci u pet dimenzija, treba iskoristiti moć estetike, uključujući boju, veličinu i podelu pogleda. Podela pogleda je definitivno jedna od najmoćnijih komponenti za izgradnju efikasne vizualizacije podataka, kao što je prikazano u donjoj vizualizaciji, gde se jasno vidi da su računari podeljeni po tipu (slika 5).



Slika 5. *Prikaz cene računara po kompanijama uz dodatne tri dimenziju predstavljene bojom. veličinom i podelom pogleda*

3.5. Pojam vremena

Neminovno postaje sve teže naći put oko ograničenja dvodimenzionalnog uređaja za prikaz da bi se vizualizovalo više dimenzija podataka. Jedan od metoda je korišćenje više aspekata odnosno podele pogleda. Osim toga, može se koristiti i pojam vremena ako skup podataka ima vremenski aspekt. U tom slučaju generisane vizuallizacije posledično mogu dati znatno poboljšan uvid u podatke i naknadnu analizu nad posmatranim podacima.

4. ZAKLJUČAK

Grafika daje kvalitativan osećaj za podatke, pomažući da se shvati šta se sa njima dešava. Dva očigledna nedostatka slojevitosti gramatike: statična je, bez interakcije i daje skromniji uvid u oblasti grafike za kategorijalne podatke. Postoji zainteresovanost za razvoj drugih okvira koji olakšavaju zajedničke zadatke u analizi podataka i izgradnji alata za grafičko zaključivanje. Interaktivna grafika je važna porodica alata jer ubrzava proces analize podataka. Sa dobro osmišljenim interaktivnim grafičkim paketom, vreme između koraka se dodatno smanjuje jer se može izmeniti prethodna reprezentacija, a ne da se počinje od nule. Zbog toga je važno proširiti slojevitu gramatiku i na opštu interakciju.

Takođe je bitan i razvoj boljih alata za druge zadatke analize podataka. Na primer, uobičajena strategija rešavanja problema je razbijanje velikog problema na male delove, rad na svakoj komponent pojedinačno, a zatim ponovno spajanje. Slojevita gramatika ne uključuje interakciju korisnika sa grafikom; svi grafici su statični i odvojeni.

Očigledno je da postoji ogroman prostor za dodavanje interakcije ovoj gramatici. Brzina je takođe izazov. Da bi se besprekorno percipirala interaktivna grafika ona se mora ažurirati više puta u sekundi. Gramatika grafike, u trenutnom obliku, moćna je i korisna, ali nije sveobuhvatna. Zato sve navedeno predstavlja izazove i potencijalne pravce za nastavak istraživanja.

5. LITERATURA

[1] Wilkinson, L. “The grammar of graphics.”, Handbook of computational statistics, Springer, 2012

[2] Towards data science, A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional Data, https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149, (pristupano 10.2021.)

[3] Mao, Yingsen. “Data visualization in exploratory data analysis: An overview of methods and technologies”. diss. 2015.

[4] Towards data science, Introduction to Plotline as the Alternative of Data Visualization Package in Python, https://towardsdatascience.com/introduction-to-plotnine-as-the-alternative-of-data-visualization-package-in-python-46011ebef7fe (pristupano 10.2021)

Kratka biografija:

Milica Damjanović rođena je 3.11.1994. godine u Somboru. Završila je gimnaziju “Gimnazija Beli Manastir” u Belom Manastiru 2013. godine. Diplomirala je na Fakultetu Tehničkih nauka u Novom Sadu 2019. godine i iste godine je upisala master studije na smeru Računarstvo i automatika.

kontakt: milicaa031@gmail.com