

Vizualizacija podatkov

V prejšnjih poglavjih smo obravnavali predvsem modeliranje podatkov in metode, s katerimi iz podatkov gradimo napovedne ali opisne modele. Pri tem smo se vizualizacije že večkrat, morda le bežno, dotaknili, na primer pri predstavitvi modelov (nomogrami), rezultatih zmanjšanja dimenzij (PCA in t-SNE) ter razlagah točkovnih grafov. A o vizualizaciji podatkov še nismo sistematično spregovorili in čas je, da to storimo, saj je vizualizacija ključna pri predstavitvi celotnega procesa in rezultatov tehnik odkrivanja znanja iz podatkov.

Vizualizacija podatkov je samostojno področje, ki raziskuje, kako podatke (in modele) predstaviti tako, da jih lahko zaznamo, razumemo in interpretiramo. Znanost grafične komunikacije podatkov (angl. *visual data communication*) povezuje računalništvo, statistiko, kognitivno psihologijo, oblikovanje in raziskave človeške percepcije. Njeno osnovno vprašanje je presenetljivo preprosto: kako podatke prikazati tako, da bomo v njih čim hitreje in čim bolj pravilno opazili vzorce, razlike, povezave in odstopanja.

Dobro zasnovana vizualizacija temelji na dejstvu, da je človekov vidni sistem izjemno zmogljiv pri zaznavanju struktur, trendov in anomalij. Pogosto lahko iz grafa skoraj v trenutku razberemo zakonitosti, ki bi jih iz tabelarnih podatkov ali statističnih povzetkov težko opazili. Vizualizacija tako ni le sredstvo za prikaz rezultatov, ampak pomembno orodje analize podatkov.

V podatkovni znanosti vizualizacijo uporabljamo v različnih fazah dela. Najprej kot orodje raziskovalne analize podatkov (angl. *exploratory analysis* in *exploratory visualization*), kjer podatkovni znanstveniki oziroma analitiki z grafičnimi prikazi preverjamo porazdelitve, iščemo osamelce, ocenjujemo povezave med atributi in odkrivamo morebitne napake v podatkih. Vizualizacija je pogosto tudi najhitrejši način za razhroščevanje podatkov, saj so napačno kodirani atributi, manjkajoče vrednosti ali nenavadne meritve na dobrem grafu pogosto takoj opazne. Pomembno vlogo ima tudi pri oblikovanju hipotez, saj nas vizualni vzorci pogosto usmerijo k vprašanjem, ki jih brez grafične predstavitve sploh ne bi zastavili. Druga pomembna vloga vizualizacije je komunikacija rezultatov, kjer je cilj poda-

Naš vidni sistem je rezultat evolucijske prilagoditve na hitro zaznavanje struktur, kontrastov, gibanja in anomalij v prostoru. Prav zato lahko grafične predstavitve podatkov pogosto razumemo bistveno hitreje kot numerične tabele.

janje naših ugotovitev, ki sledijo iz podatkov, drugim, torej ciljnemu občinstvu. Pojasnjevalna vizualizacija (*explanatory visualization*) je namenjena občinstvu ter skuša jasno in učinkovito prenesti sporočilo. Dober raziskovalni graf je lahko neurejen, poln informacij in namenjen hitremu preverjanju idej; dober pojasnjevalni graf pa mora biti jassen, osredotočen in prilagojen ciljnemu občinstvu.

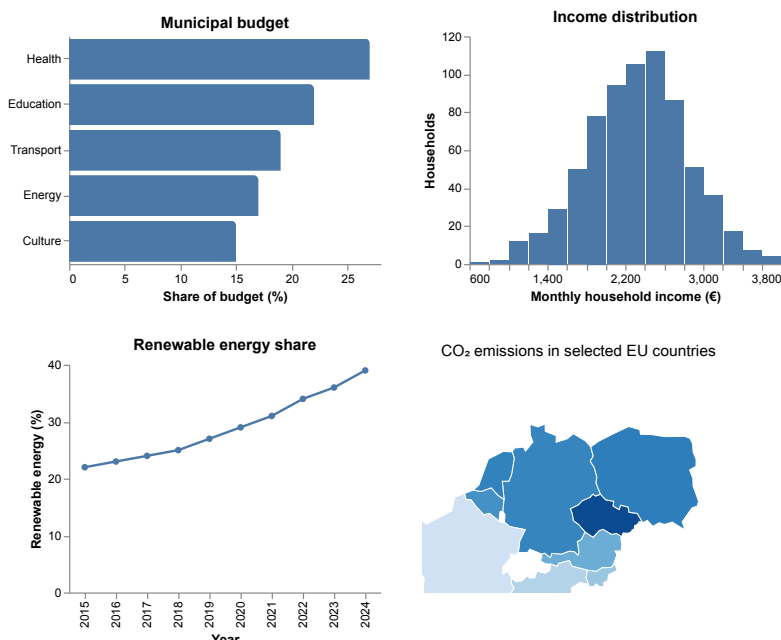
Raziskave na področju vizualizacije so pokazale, da različni grafični elementi niso enako učinkoviti. Človek na primer precej natančneje primerja položaje točk na osi kot pa površine krogov ali intenziteto barv. Zato izbira načina prikaza podatkov ni zgolj estetsko vprašanje, ampak neposredno vpliva na pravilnost interpretacije. Napačno izbrane vizualizacije lahko podatke popačijo, zavedejo opazovalca ali prikrijejo pomembne zakonitosti.

Tu bomo obravnavali osnovna načela vizualizacije podatkov in pregledali najpomembnejše tipe grafičnih prikazov. Zanimalo nas bo, kako različne vrste podatkov učinkovito predstaviti, katere vizualne elemente uporabljamo za prikaz vrednosti in odnosov med podatki ter kako izbrati prikaz, ki ustreza analitični nalogi. Posebej bomo poudarili tudi omejitve posameznih tipov grafičnih predstavitev, omenili nekaj pogostih napak pri načrtovanju vizualizacij in se na kratko posvetili interpretaciji kompleksnih vizualnih predstavitev. Učinkovite vizualizacije, predstavljene na računalnikih, podpirajo interaktivnost, zato bomo omenili tudi področje vizualne analitike (angl. *visual analytics*). Poglavje zaključimo praktično, s pregledom nekaterih programskih pristopov za konstrukcijo vizualizacij.

Tipi podatkov

Pri načrtovanju vizualizacij je najprej potrebno razumeti, kakšne podatke sploh prikazujemo. Različni tipi podatkov zahtevajo različne tipe grafičnih predstavitev, prav tako pa niso vsi vizualni elementi enako učinkoviti pri prikazu posameznih vrst podatkov. Prav zaradi pravilne (ali napačne) uporabe vizualnih dokumentov so lahko nekateri grafi pregledni in informativni, drugi pa zavajajoči ali težko razumljivi.

V podatkovni znanosti običajno ločimo med kategoričnimi, numeričnimi, časovnimi in prostorskimi podatki. Kategorični podatki opisujejo pripadnost skupinam ali razredom. Primeri so spol, vrsta izdelka, država ali oznaka razreda. Pri takih podatkih nas pogosto zanimajo frekvence, deleži ali primerjave med skupinami. Numerični podatki predstavljajo merljive količine, kot so temperatura, dohodek, masa ali starost. Pri njih nas zanimajo porazdelitve, povprečja, korelacije in trendi. Posebna vrsta numeričnih podatkov so časovni podatki, kjer vrednosti opazujemo skozi čas. Takšne podatke pogosto



Slika 33: Primeri različnih tipov podatkov: kategorični podatki (stolpčni graf), numerični podatki (histogram), časovni podatki (linijski graf) in prostorski podatki (zemljevid).

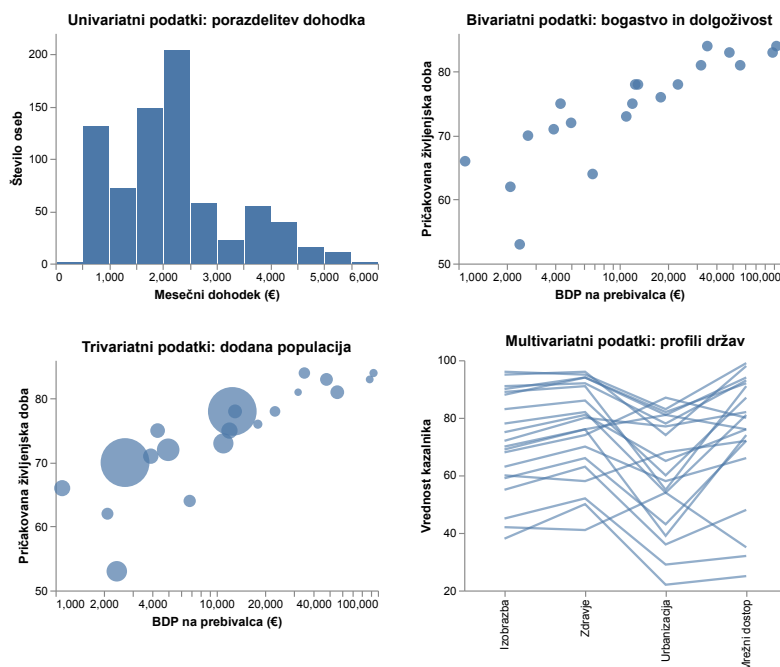
predstavljamo z linijskimi grafi, saj želimo poudariti razvoj ali spremembo skozi čas. Prostorski podatki pa vključujejo geografsko ali drugo prostorsko komponento, zato jih pogosto prikazujemo na zemljevidih ali prostorskih mrežah.

Slika 33 prikazuje tipične primere vizualizacij za različne vrste podatkov. Že iz teh primerov je razvidno, da izbira grafične predstavitve ni naključna. Linijski graf je primeren za prikaz časovnih sprememb, histogram za prikaz porazdelitve numeričnih podatkov, zemljevid pa za prikaz prostorske razporeditve.

Pomembna je tudi dimenzionalnost podatkov. Pri enorazsežnih oziroma univariatnih podatkih (*univariate data*) analiziramo eno samo spremenljivko. Zanima nas na primer porazdelitev starosti ali višine oseb. Pri dvodimenzionalnih oziroma bivariatnih podatkih proučujemo odnos med dvema spremenljivkama, na primer povezavo med višino in telesno maso. Večrazsežni oziroma multivariatni podatki vsebujejo več atributov hkrati in zahtevajo bolj kompleksne vizualizacije.

Slika 34 prikazuje vizualizacije prilagojene različnim dimenzionalnostim podatkov. Histogram omogoča vpogled v porazdelitev ene same spremenljivke, razsevni diagram prikazuje povezavo med dvema spremenljivkama, paralelne koordinate pa omogočajo prikaz več atributov hkrati. Z naraščanjem števila dimenzij postaja interpretacija vizualizacij vse zahtevnejša, zato je pomembno izbrati prikaz, ki omogoča preglednost tudi pri večji kompleksnosti podatkov.

Prikaz s paralelnimi koordinatami hitro postane nepregleden pri večjem številu primerov ali atributov, saj se črte prekrivajo in otežijo zaznavanje vzorcev ter povezav med spremenljivkami, sama vizualizacija in nje interpretacija pa postane močno odvisna od vrstnega reda atributov.



Slika 34: Grafični prikazi za različne dimenzionalnosti vizualiziranih podatkov: histogram za univariatne podatke, razsewni diagram za bivariatne podatke, mehurčni diagram za trivariatne podatke in paralelne koordinate za multivariatne podatke.

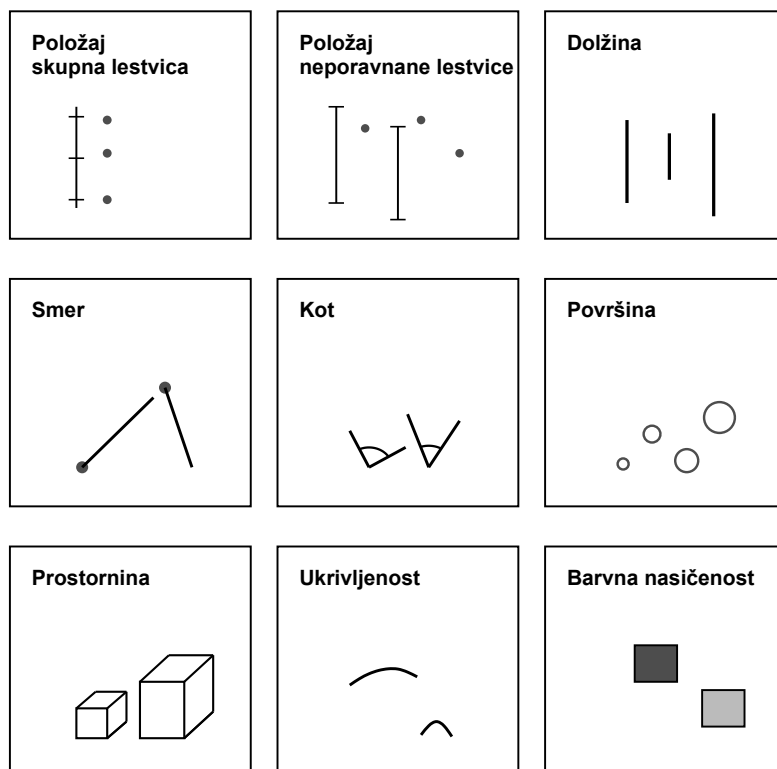
Vizualni elementi

Vizualizacije temeljijo na preslikavi podatkov v vizualne elemente oziroma vizualne kanale. Ti določajo, kako bodo podatki predstavljeni opazovalcu. Najpomembnejši vizualni elementi so položaj, dolžina, kot, površina, barva, oblika in velikost.

Položaj je najpomembnejši in hkrati najnatančnejši vizualni element. Človek zelo dobro zaznava razlike v položaju točk na skupni osi, zato so razsewni diagrami, linijski grafi in stolpčni grafi pogosto zelo učinkoviti. Dolžina je prav tako zelo dobro zaznavna in jo uporabljamo predvsem pri stolpčnih grafih. Precej manj natančno zaznavamo kote in površine, zato so tortni diagrami pogosto slabša izbira od stolpčnih grafov. Barva je izredno uporabna za ločevanje skupin, označevanje intenzitete ali poudarjanje pomembnih delov podatkov, vendar človek razlik v intenziteti barv ne zaznava posebej natančno. Oblika in velikost sta predvsem pomožna vizualna elementa za razlikovanje kategorij ali poudarjanje posameznih primerov.

Na sliki 35 so prikazani nekateri pogosti vizualni elementi. Pomembno je razumeti, da različni elementi niso enako učinkoviti. Položaj in dolžina omogočata zelo natančne primerjave med vrednostmi, medtem ko primerjava površin ali kotov pogosto vodi do napačnih ocen (tabela 11). Prav zato so nekateri tipi grafov bolj primerni za analitično delo kot drugi.

Pri načrtovanju vizualizacij moramo zato razumeti tako podatke



Slika 35: Najpogostejši vizualni elementi za predstavitev podatkov: položaj, dolžina, kot, površina, barva, oblika in velikost. Vizualni elementi se razlikujejo po natančnosti, s katero človek zaznava kvantitativne razlike.

kot tudi način, kako ljudje zaznavamo vizualne elemente. Preden izberemo grafični prikaz, moramo vedeti, kakšne podatke imamo, kakšne odnose želimo prikazati in kako človek zaznava posamezne vizualne elemente. Šele nato lahko izberemo vizualizacijo, ki bo podatke predstavila jasno, učinkovito in brez zavajanja.

Osnovni tipi grafov

Izbira vizualizacije je seveda odvisna od vprašanja, na katerega želimo odgovoriti s podatki. Različni grafi poudarjajo različne lastnosti podatkov: porazdelitve, trende, povezave, negotovost ali strukturo. Univerzalno najboljši graf ne obstaja, obstajajo pa dobre in slabe vizualizacije. Vsak tip grafa je primeren za določen analitični cilj in lahko zavaja, če ga uporabimo v neprimernem kontekstu, po drugi strani pa ga lahko uporabimo tako, da primerno izpostavi idejo ali vzorec, zaradi katerega smo se odločili podatke grafično prikazati.

V nadaljevanju predstavimo nekaj osnovnih tipov grafov. Pri vsakem tipu grafa bomo obravnavali:

- na kakšno vprašanje odgovarja,
- kakšne so njegove prednosti in omejitve,

Vizualni element	Tipična uporaba	Primeri	Zaznavna natančnost
Položaj	Primerjava vrednosti	Razsevni diagrami	Zelo visoka
Dolžina	Primerjava količin	Stolpčni grafi	Visoka
Kot	Prikaz deležev	Tortni diagrami	Srednja
Velikost	Poudarjanje pomembnosti	Mehurčni diagrami	Srednja
Orientacija	Smer ali usmerjenost	Vektorska polja	Srednja
Oblika	Razlikovanje kategorij	Označevalci točk	Omejena
Tekstura	Razlikovanje področij	Kartografija	Omejena
Ukrivljenost	Prikaz trendov ali tokov	Diagrami povezav	Omejena
Barva	Skupine ali intenziteta	Toplotni zemljevidi	Nizka
Svetlost	Intenziteta ali gostota	Sivinske skale	Nizka
Površina	Velikost količine	Mehurčni diagrami	Nizka
Volumen	3D prikazi količin	3D stolpci	Zelo nizka

- katere so tipične pogoste napake pri njegovi uporabi ali interpretaciji.

Histogram. Histogram (slika 36) prikazuje porazdelitev numerične spremenljivke tako, da razdeli območje vrednosti v intervale (koše) in prikaže število primerov v posameznem intervalu.

Histogrami odgovarjajo na vprašanja:

- Ali je porazdelitev simetrična ali asimetrična?
- Ali vsebuje več vrhov?
- Ali so prisotni osamelci?
- Ali je porazdelitev približno normalna?

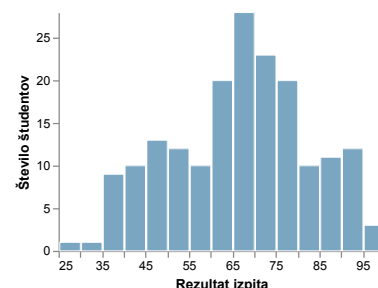
Histogrami omogočajo hiter vpogled v obliko porazdelitve podatkov. Primerni so za raziskovalno analizo in hitro razumevanje oblike podatkov. Pri histogramih rezultat močno vpliva izbira širine intervalov. Preširoki intervali skrijejo strukturo, preozki pa poudarijo šum v podatkih.

Pogosta napaka je primerjanje histogramov z različno velikimi intervali ali različnimi velikostmi vzorca brez ustrezne normalizacije. Histogrami tudi niso primerni za podatke z majhnim številom primerov, na primer, kadar imamo le nekaj deset opazovanj, saj lahko oblika porazdelitve tedaj močno zavisi od izbire intervalov in naključnih odstopanj v podatkih.

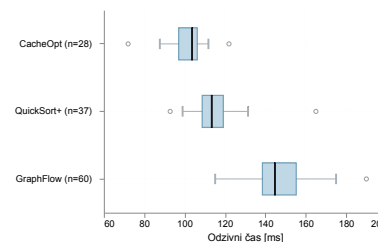
Škatla z brki. Škatla z brki (angl. *box-and-whisker plot*, slika 37) povzema porazdelitev numerične spremenljivke s kvartili in osamelci. Osamelci so podatkovne točke, ki ležijo več kot 1,5 medkvartilnega razmika nad tretjim kvartilom ali pod prvim kvartilom.

Odgovarja na vprašanja:

Tabela 11: Vizualni elementi za predstavitev podatkov in njihove lastnosti.



Slika 36: Histogram porazdelitve rezultatov izpita. Graf razkrije asimetrijo in morebitno večvršnost porazdelitve.



Slika 37: Škatlasti diagram z brki prikazuje porazdelitev odzivnih časov treh algoritmov. Črna črta v škatli označuje mediano, robova škatle predstavljata prvi in tretji kvartil, brki segajo do najbolj oddaljenih vrednosti brez osamelcev, krogi pa označujejo osamelce.

- Kakšna je mediana podatkov?
- Kako velika je variabilnost?
- Ali so prisotni osamelci?
- Kako se porazdelitve razlikujejo med skupinami?

Škatlasti diagrami omogočajo kompaktno primerjavo več skupin hkrati. Pri tem pa izgubimo podrobnejši vpogled v obliko porazdelitve. Dve zelo različni porazdelitvi lahko ustvarita skoraj enako vizualizacijo. Pogosta napaka je interpretacija osamelcev kot napak merjenja. Osamelci lahko predstavljajo pomembne redke dogodke. Škatle z brki niso primerne za zelo majhne vzorce.

Violinski diagram. Violinski diagram (angl. *violin plot*, slika 38) razširi škatlasti diagram z brki z oceno gostote porazdelitve podatkov. Širina violine predstavlja gostoto podatkov pri posamezni vrednosti: širši deli označujejo območja z več opazovanji. V prikazu so posamezna opazovanja prikazana s točkami, črna črta pa označuje mediano.

Odgovarja na enaka vprašanja kot škatla z brki, dodatno pa razkrije:

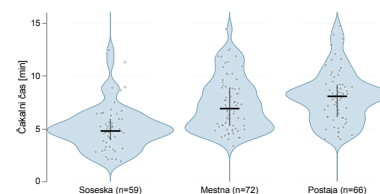
- večvršnost porazdelitve,
- asimetrijo,
- podrobnejšo strukturo podatkov,
- območja z večjo koncentracijo opazovanj.

Violinski diagrami prikažejo tudi gostoto in strukturo porazdelitve. Prikaz je občutljiv na izbiro metode glajenja pri oceni gostote. Navidezni vrhovi ali doline so lahko posledica izbire parametrov glajenja in ne dejanske strukture podatkov. Pogosta napaka je uporaba violinskih diagramov pri zelo majhnih vzorcih, kjer ocena gostote ni zanesljiva.

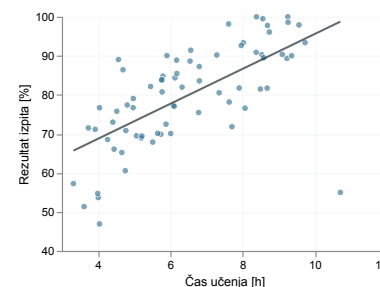
Razsevni diagram. Razsevni diagram (angl. *scatter plot*, slika 39) prikazuje odnos med dvema numeričnima spremenljivkama. Vsaka točka predstavlja eno opazovanje, njen položaj pa določata vrednosti obeh spremenljivk. Diagram omogoča vizualno prepoznavanje povezav, trendov, gruč in osamelcev.

Odgovarja na vprašanja:

- Ali sta spremenljivki povezani?
- Ali je povezava linearna ali nelinearna?
- Ali obstajajo gručice ali podskupine?



Slika 38: Violinski diagram prikazuje porazdelitve čakalnih časov v različnih trgovinah. Širina violine predstavlja gostoto podatkov, točke posamezna opazovanja, črna črta pa mediano.



Slika 39: Razsevni diagram prikazuje povezavo med časom učenja in rezultatom izpita. Vsaka točka predstavlja posameznega študenta, črta pa linearni trend med obema spremenljivkama. Opazen je tudi osamelec, ki odstopa od splošnega vzorca podatkov.

- Ali so prisotni osamelci?
- Kako močna je povezanost med spremenljivkama?

Razsevni diagram neposredno pokaže odnos med spremenljivkama. Pogosto sodijo med najpomembnejša orodja raziskovalne analize podatkov, saj omogočajo hitro prepoznavanje trendov, vzorcev in nepravilnosti. Njihova glavna omejitev je prekrivanje točk pri velikih podatkovnih množicah, kjer gosta območja postanejo težko berljiva. Takrat si pogosto pomagamo s prosojnostjo točk, vzorčenjem ali alternativnimi prikazi gostote.

Pogosta napaka pri interpretaciji razsevnih diagramov je sklepanje o vzročnosti na podlagi korelacije. Diagram prikazuje povezanost med spremenljivkama, ne pa nujno vzročnega odnosa.

Črtni diagram. Črtni diagram (angl. *line chart*, slika 40) prikazuje spremembe vrednosti vzdolž urejene osi, najpogosteje skozi čas. Posamezne meritve so predstavljene s točkami, ki so povezane s črto, kar omogoča vizualno spremljanje trendov, nihanj in sprememb skozi čas.

Odgovarja na vprašanja:

- Kako se količina spreminja skozi čas?
- Ali obstajajo trendi ali sezonski vzorci?
- Ali so prisotne nenadne spremembe?
- Kako hitro ali pogosto se vrednosti spreminjajo?

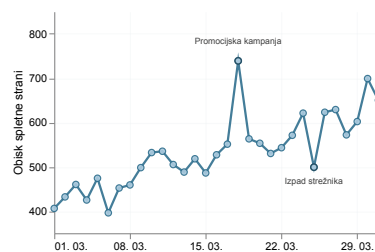
Črtni diagrami so posebej primerni za prikaz časovnih trendov in sprememb. Posebej primerni so za prikaz časovnih vrst, kjer želimo poudariti razvoj podatkov skozi čas.

Njihova slabost je implicitna predpostavka zveznosti med zaporednimi točkami, ki ni vedno smiselna. Pri velikem številu serij ali zelo gostih meritvah lahko graf hitro postane nepregleden. Pogosta napaka pri uporabi tega grafa je povezovanje kategoričnih podatkov s črto, kar ustvarja vtis neobstoječe zveznosti. Težava je tudi prikaz preveč podatkovnih serij hkrati, saj to oteži primerjavo in interpretacijo.

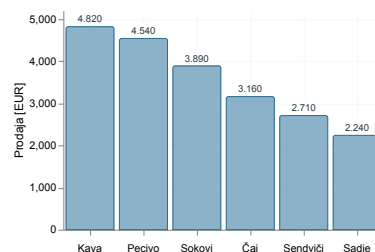
Stolpčni diagram. Stolpčni diagram (angl. *bar chart*, slika 41) primerja količine med kategorijami. Dolžina oziroma višina stolpcev predstavlja vrednost posamezne kategorije, kar omogoča hitro vizualno primerjavo med skupinami.

Odgovarja na vprašanja:

- Katera kategorija ima največjo ali najmanjšo vrednost?



Slika 40: Črtni diagram prikazuje obisk spletne strani skozi čas. Črta povezuje zaporedne meritve in omogoča prepoznavanje trendov ter nenadnih sprememb. Označena sta tudi dva posebna dogodka oziroma vpliv akcij.



Slika 41: Stolpčni diagram primerja prodajo med kategorijami izdelkov. Višina stolpcev predstavlja prodajo posamezne kategorije.

- Kako velike so razlike med skupinami?
- Kakšen je vrstni red kategorij?
- Katere kategorije izstopajo?

Stolpčne diagrame običajno hitro in intuitivno razumemo. Posebej primerni so za predstavitev agregiranih podatkov in primerjanje kategorij širšemu občinstvu.

Njihova omejitev je, da prikazujejo predvsem povzetke podatkov in skrijejo variabilnost znotraj skupin. Pri večjem številu kategorij lahko diagram hitro postane nepregleden. Pogosta napaka je rezanje navpične osi, saj uporabniki primerjajo dolžine stolpcev. Že majhna sprememba začetka osi lahko močno pretirava ali zmanjša zaznane razlike med kategorijami.

Toplotna karta. Toplotna karta (angl. *heatmap*, slika 42) prikazuje vrednosti matrice s pomočjo barvne lestvice. Posamezne celice predstavljajo vrednosti med pari spremenljivk, pri čemer barva označuje velikost ali intenzivnost vrednosti. Toplotne karte omogočajo hitro prepoznavanje vzorcev, povezanosti, gruč in območij z izrazitimi vrednostmi.

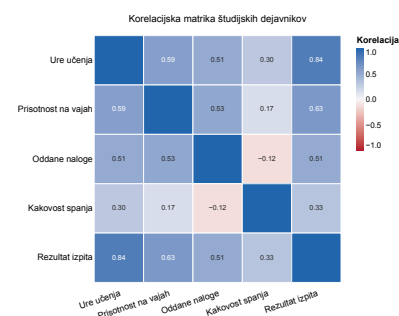
Odgovarja na vprašanja:

- Katere spremenljivke so močno povezane?
- Ali obstajajo vzorci ali gruče?
- Kje so skoncentrirane visoke ali nizke vrednosti?
- Katere spremenljivke imajo podobno vedenje?

Toplotne karte omogočajo hkraten pregled velikega števila vrednosti in povezav med njimi. Posebej uporabne so pri analizi korelacijskih matrik, časovnih vzorcev in drugih strukturiranih podatkov.

Njihova slabost je odvisnost od izbire barvne lestvice in razvrstitve podatkov. Slabo izbrane barve lahko popačijo interpretacijo razlik med vrednostmi ali poudarijo navidezne vzorce. Pri podatkih z naravno sredinsko vrednostjo, kot so korelacije, je smiselna uporaba divergentnih barvnih lestvic z nevtralno sredinsko barvo, na primer modro–belo–rdeče lestvice.

Pogosta napaka je uporaba mavričnih barvnih lestvic, ki ustvarijo umetne kontraste in otežijo interpretacijo. Priporočljive so perceptualno enakomerne barvne lestvice, kjer spremembe v barvi ustrezajo spremembam v podatkih. Za sekvenčne podatke so pogosto primerne lestvice, kot je *viridis*, za korelacijske matrice pa divergentne lestvice, kot sta *coolwarm* ali *RdBu*. Pomembna je tudi zmerna nasičenost barv in dobra čitljivost pri sivinskem prikazu.



Slika 42: Toplotna karta prikazuje korelacijsko matriko med številom ur učenja, prisotnostjo na vajah, številom oddanih nalog, kakovostjo spanja in rezultatom izpita. Vsaka celica predstavlja korelacijo med parom spremenljivk: temno modra označuje močno pozitivno korelacijo, temno rdeča močno negativno korelacijo, svetlejši odtenki pa šibkejšo povezanost. Diagonala vsebuje popolne korelacije spremenljivk s samimi seboj.

Matrika razsevnih diagramov. Matrika razsevnih diagramov (angl. *scatterplot matrix* ali *pair plot*, slika 43) prikazuje vse parne odnose med več numeričnimi spremenljivkami hkrati. Vsaka celica matrike vsebuje razsevni diagram za izbrani par spremenljivk, diagonala pa pogosto prikazuje porazdelitve posameznih spremenljivk.

Odgovarja na vprašanja:

- Katere spremenljivke so povezane?
- Ali obstajajo gručice ali ločeni razredi?
- Katere spremenljivke so redundantne?
- Ali so povezave linearne ali nelinearne?
- Ali so prisotni osamelci ali nenavadni vzorci?

Matrike razsevnih diagramov omogočajo hiter pregled odnosov med več spremenljivkami. Omogočajo istočasno prepoznavanje korelacij, gruč, osamelcev in morebitnih nelinearnih povezav.

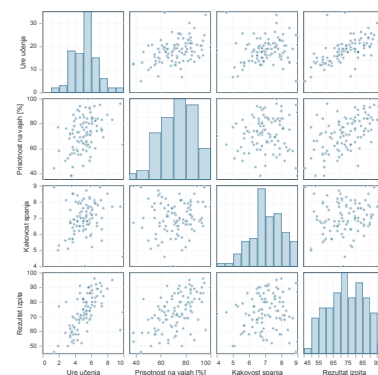
Njihova glavna slabost je slaba skalabilnost. Število grafov raste kvadratno s številom spremenljivk, zato matrike hitro postanejo nepregledne. Pri velikem številu primerov se pojavi tudi prekrivanje točk, kar oteži interpretacijo.

Pogosta napaka je uporaba matrik razsevnih diagramov pri zelo velikem številu spremenljivk ali opazovanj, kjer vizualizacija izgubi preglednost in informativnost.

Vizualizacija pri raziskovalni analizi podatkov

Vizualizacija ima pomembno vlogo pri raziskovalni analizi podatkov (angl. *exploratory data analysis*, EDA). Preden zgradimo modele ali izvedemo zahtevnejše analize, moramo podatke najprej razumeti. Vizualizacija nam omogoča hiter vpogled v strukturo podatkov, kakovost meritev, povezave med atributi in morebitne nenavadne vzorce. Pogosto lahko že preprost graf razkrije zakonitosti, ki jih iz tabelarnih podatkov ali numeričnih povzetkov težko opazimo. Pri tem nas na primer zanimajo naslednja vprašanja:

- **Kakšna je porazdelitev atributov?** Za pregled porazdelitev numeričnih atributov uporabljamo predvsem histograme, gostotne diagrame in violinske diagrame. Takšni prikazi nam omogočajo oceno razpršenosti podatkov, večvršnosti porazdelitve ter prisotnosti dolgih repov. Pri kategoričnih podatkih pogosto uporabimo stolpčne diagrame.
- **Ali so podatki simetrični ali asimetrični?** Asimetrijo porazdelitve lahko hitro opazimo na histogramih ali gostotnih diagramih.



Slika 43: Matrika razsevnih diagramov prikazuje parne odnose med številom ur učenja, prisotnostjo na vajah, kakovostjo spanja in rezultatom izpita. Vsaka celica vsebuje razsevni diagram za določen par spremenljivk, diagonala pa prikazuje porazdelitve posameznih spremenljivk. Vidna je pozitivna povezava med časom učenja in rezultatom izpita ter šibkejša povezanost drugih dejavnikov.

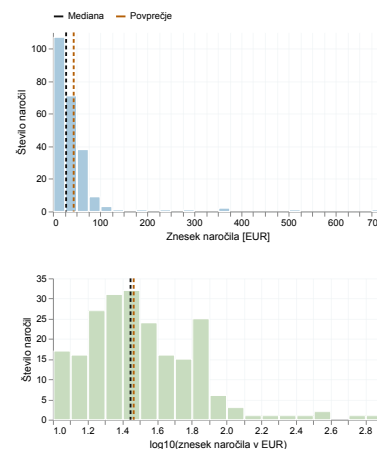
Močno desno asimetrične porazdelitve so pogoste pri finančnih podatkih, številu ogledov spletnih vsebin ali velikosti datotek.

- **Ali so prisotni osamelci oziroma anomalije?** Osamelce pogosto odkrivamo s škatlami z brki, razsevnimi diagrami ali časovnimi prikazi. Nenavadne točke lahko predstavljajo napake meritev ali pomembne redke dogodke, kot so goljufive transakcije ali okvare sistemov.
- **Ali v podatkih manjkajo vrednosti?** Manjkajoče vrednosti lahko prikažemo s posebnimi matrikami prisotnosti podatkov ali s toplotnimi kartami. Takšni prikazi pogosto razkrijejo sistemske težave pri zajemu podatkov ali manjkajoče meritve v določenih časovnih obdobjih.
- **Katere spremenljivke so med seboj povezane?** Povezave med numeričnimi atributi najpogosteje raziskujemo z razsevnimi diagrami, matrikami razsevnih diagramov in korelacijskimi toplotnimi kartami. Takšni prikazi lahko razkrijejo linearne ali nelinearne povezave ter redundantne attribute.
- **Ali obstajajo gručice ali podskupine primerov?** Gruče podobnih primerov lahko pogosto opazimo na razsevnih diagramih ali projekcijah večdimenzionalnih podatkov v ravnino. Takšni prikazi so pomembni pri segmentaciji uporabnikov, analizi dokumentov in bioinformatiki.
- **Ali podatki vsebujejo trende ali sezonske vzorce?** Pri časovnih podatkih uporabljamo predvsem črtne diagrame in časovne toplotne karte. Ti omogočajo zaznavanje dolgoročnih trendov, sezonskih vzorcev, periodičnosti in nenadnih sprememb v podatkih.

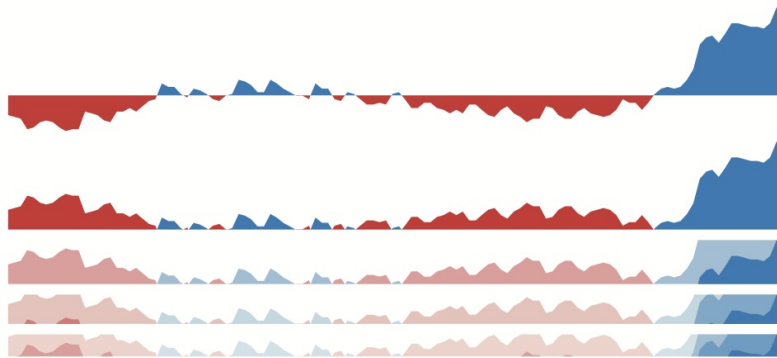
Raziskovalna analiza podatkov ni le uvodni korak projekta, ampak spremlja skoraj celoten proces podatkovnega rudarjenja. Vizualizacijo uporabljamo pri preverjanju kakovosti podatkov, izbiri atributov, razumevanju modelov in interpretaciji rezultatov. Prav zato je vizualizacija eno ključnih orodij sodobne podatkovne znanosti.

Nekaj primerov kompleksnih vizualizacij

V prejšnjih razdelkih smo obravnavali predvsem osnovne tipe grafičnih prikazov, ki jih pogosto uporabljamo pri raziskovalni analizi podatkov in predstavitvi rezultatov. Mnogi resnični podatkovni problemi pa vključujejo bolj kompleksne strukture: hierarhije, omrežja, prostorske tokove ali zelo goste časovne podatke. Pri takih podatkih osnovni grafi pogosto niso več dovolj učinkoviti, zato uporabljamo



Slika 44: Porazdelitev zneskov spletnih naročil je izrazito desno asimetrična: večina naročil ima nizke zneske, nekaj zelo velikih naročil pa tvori dolgi desni rep porazdelitve. Zaradi teh ekstremnih vrednosti je povprečje večje od mediane. Spodaj je prikazana ista porazdelitev po logaritemski transformaciji, ki zmanjša vpliv velikih vrednosti in porazdelitev naredi bolj simetrično ter primernejšo za nadaljnjo analizo in modeliranje.



naprednejše vizualizacijske pristope, ki skušajo v omejenem prostoru prikazati več informacij hkrati.

Kompleksne vizualizacije praviloma niso namenjene zgolj estetskemu učinku. Večina jih nastane kot odgovor na konkreten problem predstavitve podatkov: kako prikazati veliko število časovnih vrst, kako predstaviti hierarhične odnose med objekti, kako razumljivo prikazati omrežje povezav ali kako hkrati prikazati geografski položaj, smer gibanja in količino podatkov. Takšne vizualizacije pogosto zahtevajo nekoliko več časa za razumevanje, vendar lahko po začetnem učenju omogočajo bistveno učinkovitejšo analizo kompleksnih podatkovnih struktur.

Horizontalni graf. Horizontalni graf (angl. *horizon graph*, slika 45) je poseben način prikaza časovnih vrst, namenjen predvsem povečanju gostote prikazanih podatkov. Klasični črtni ali površinski graf pri večjem številu časovnih vrst hitro postane nepregleden, saj posamezna serija potrebuje precej navpičnega prostora. Horizontalni graf skuša ta problem rešiti tako, da podatke razdeli v več pasov in jih nato zloži enega čez drugega. Osnovna ideja je, da negativne vrednosti zrcalimo nad osnovno os, nato pa višinske pasove zložimo enega čez drugega. S tem močno zmanjšamo potrebno višino grafa, pri čemer ohranimo časovno ločljivost podatkov. Takšni prikazi so posebej uporabni pri nadzornih ploščah in sistemih za spremljanje velikega števila časovnih signalov. Prednost horizontalnih grafov je izjemna prostorska učinkovitost. V istem prostoru lahko prikažemo bistveno več podatkov kot s klasičnimi črtnimi diagrami. Slabost pa je nekoliko težje začetno razumevanje prikaza, saj mora opazovalec razumeti zlaganje pasov in pomen barvnih nivojev. Horizontalni grafi zato niso najprimernejši za širše občinstvo, zelo uporabni pa so pri analitičnem delu strokovnjakov.

Tokovni zemljevid. Tokovni zemljevid (angl. *flow map*) prikazuje gibanje količin skozi prostor. Takšne vizualizacije pogosto uporabl-

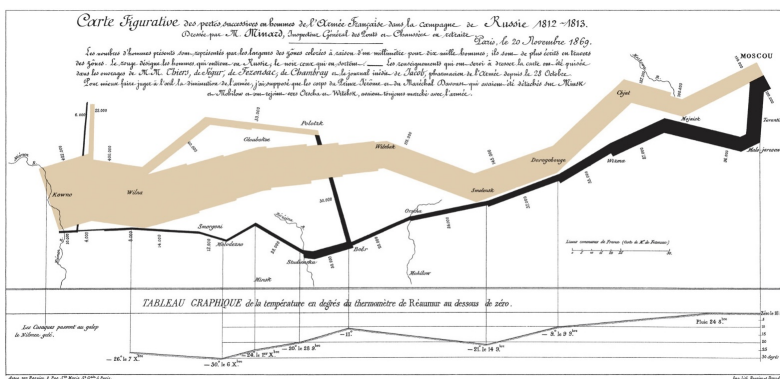
Slika 45: Horizontalni graf prikazuje časovno vrsto brez-poselnosti skozi čas. Pozitivne in negativne vrednosti so prikazane z različnimi barvami, vrednosti pa so razdeljene v več pasov, ki se prekrivajo. Slika prikazuje tudi postopno transformacijo vizualizacije: od običajnega površinskega grafa, prek zrcaljenja negativnih vrednosti nad osnovno os, do razdelitve podatkov v več nivojev, ki so nato zloženi drug čez drugega. S tem se bistveno poveča gostota prikazanih podatkov, saj lahko enako časovno ločljivost prikažemo na precej manjši površini. Povzeto po Heer in sod. (2010).

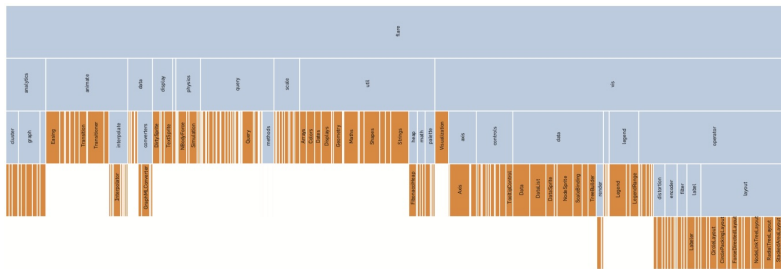
jamo za prikaz migracij, trgovinskih tokov, transporta, širjenja bolezni ali zgodovinskih premikov vojska. Slika 46 prikazuje znameniti Minardov prikaz Napoleonovega pohoda na Moskvo. Vizualizacija velja za eno najboljših predstavitev podatkov vseh časov, saj v enem samem grafu združuje več različnih dimenzij podatkov. Debelina toka predstavlja velikost vojske, smer toka prikazuje gibanje, položaj predstavlja geografijo, dodatni graf pod zemljevidom pa prikazuje temperaturo med umikom vojske. Tokovni zemljevidi so zanimivi predvsem zato, ker hkrati združujejo prostorske, časovne in količinske informacije. Dobro zasnovan tokovni zemljevid omogoča zelo intuitivno razumevanje kompleksnih procesov gibanja skozi prostor. Njihova slabost pa je možnost velike vizualne nepreglednosti pri večjem številu tokov, saj se poti hitro prekrivajo. Zato pogosto uporabljamo poenostavljanje poti, prosojnost ali interaktivno filtriranje.

Drevesni zemljevid. Drevesni zemljevid (angl. *treemap*) je vizualizacija hierarhičnih podatkov, kjer so elementi predstavljeni z vgnezenimi pravokotniki. Površina posameznega pravokotnika običajno predstavlja velikost ali pomembnost elementa, hierarhija pa je prikazana z vgnezenostjo območij. Na sliki 47 je prikazan drevesni zemljevid hierarhije programskih paketov. Večji pravokotniki predstavljajo večje oziroma pomembnejše dele sistema. Ker drevesni zemljevid učinkovito izkorišča prostor, lahko na relativno majhni površini prikažemo zelo velike hierarhične strukture.

Drevesne vizualizacije pogosto uporabljamo za prikaz uporabe diskovnega prostora, strukture finančnih trgov, organizacije datotek ali hierarhičnih klasifikacij. Ti prikazi so posebej uporabni, kadar želimo hitro oceniti relativne velikosti posameznih delov sistema. Njihova prednost je zelo učinkovita izraba prostora in dobra podpora primerjanju velikosti, slabost pa je nekoliko težje sledenje globlji hierarhični strukturi, saj pri zelo velikem številu nivojev vgnezenost

Slika 46: Tokovni zemljevid Napoleonovega pohoda na Moskvo prikazuje gibanje vojske skozi prostor in čas. Debelina toka predstavlja velikost vojske, položaj označuje geografsko lokacijo, spodnji graf pa prikazuje temperaturo med umikom. Vizualizacija velja za eno najbolj znanih predstavitev večdimenzionalnih podatkov, ki jo je leta 1869 objavil francoski inženir Charles Joseph Minard v delu *Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813*.





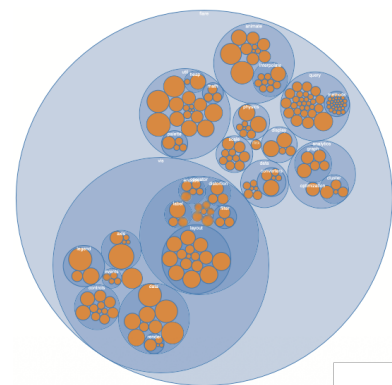
postane nepregledna. Prav tako primerjanje zelo podolgovatih pravokotnikov ni vedno intuitivno.

Sončni izsek. Sončni izsek (angl. *sunburst diagram*) je radialna različica prostorsko zapolnjujoče predstavitve hierarhij. Hierarhija je prikazana s koncentričnimi krožnimi pasovi, kjer notranji krogi predstavljajo višje nivoje hierarhije, zunanji pa nižje nivoje. Na sliki 48 je prikazan sončni izsek iste hierarhične strukture kot pri drevesnem zemljevidu. V primerjavi z drevesnim zemljevidom sončni izsek pogosto bolj jasno poudari globino hierarhije, saj se nivoji naravno širijo navzven iz središča.

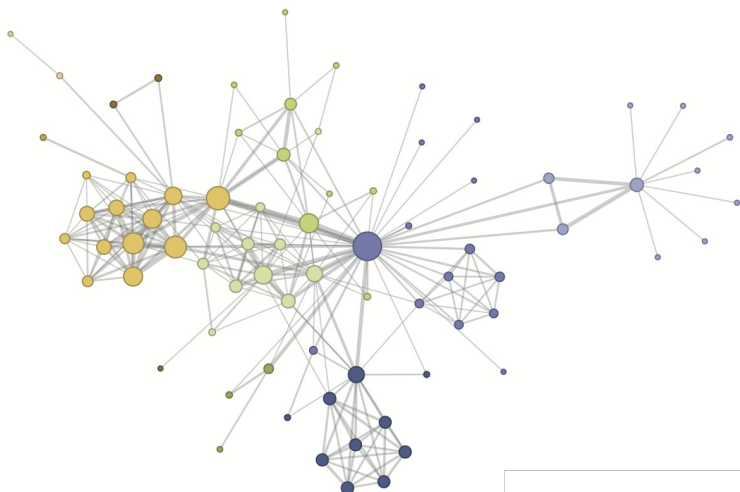
Vizualizacije tega tipa so lahko estetsko zelo privlačne in pogosto uporabljene v interaktivnih sistemih za raziskovanje podatkov. Posebej primerne so za prikaz hierarhičnih struktur, kjer nas zanima predvsem organizacija podatkov in odnosi med nivoji hierarhije. Njihova slabost je nekoliko slabša natančnost pri primerjanju površin in kotov, posebej pri zunanjih segmentih diagrama. Pri zelo globokih hierarhijah postanejo zunanji deli hitro preozki za učinkovito prikazovanje oznak.

Vizualizacija omrežij s silami. Veliko podatkovnih struktur lahko predstavimo kot omrežje povezav med objekti. Primeri vključujejo socialna omrežja, povezave med spletnimi stranmi, citatne mreže znanstvenih člankov ali biološke interakcije med proteini. Eden najpogostejših pristopov za prikaz takšnih podatkov so razporeditve na osnovi sil (angl. *force-directed layouts*). Na sliki 49 je prikazan primer omrežja likov iz romana *Les Misérables*. Vozlišča predstavljajo posamezne like, povezave pa njihovo skupno pojavljanje v besedilu. Razporeditev vozlišč nastane s simulacijo fizikalnega sistema: povezani elementi se privlačijo, nepovezani pa odbijajo. Posledica je organska razporeditev, kjer se močno povezane skupine naravno združijo v gručice. Vizualizacije tovrstnih grafov lahko omogočajo hitro prepoznavanje skupin, osrednjih vozlišč in mostov med različnimi deli omrežja. So med najpomembnejšimi pristopi pri analizi socialnih omrežij in kompleksnih grafov. Njihova glavna težava je

Slika 47: Drevesni zemljevid prikazuje hierarhijo programskih paketov. Posamezni pravokotniki predstavljajo razrede oziroma module, njihova površina pa velikost ali pomembnost posameznega elementa. Povzeto po Heer, Bostock in Ogievetsky, Figure 4f.



Slika 48: Sončni izsek prikazuje hierarhijo programskih paketov v radialni obliki. Notranji krogi predstavljajo višje nivoje hierarhije, zunanji pa podrobnejše podstrukture. Povzeto po Heer, Bostock in Ogievetsky, Figure 4e.



Slika 49: Vizualizacija omrežja z razporeditvijo na osnovi sil. Vozlišča predstavljajo literarne like, povezave pa njihove skupne pojavitve v poglavjih romana. Položaj vozlišč je določen s simulacijo privlačnih in odbojnih sil. Povzeto po Heer, Bostock in Ogievetsky, Figure 5a.

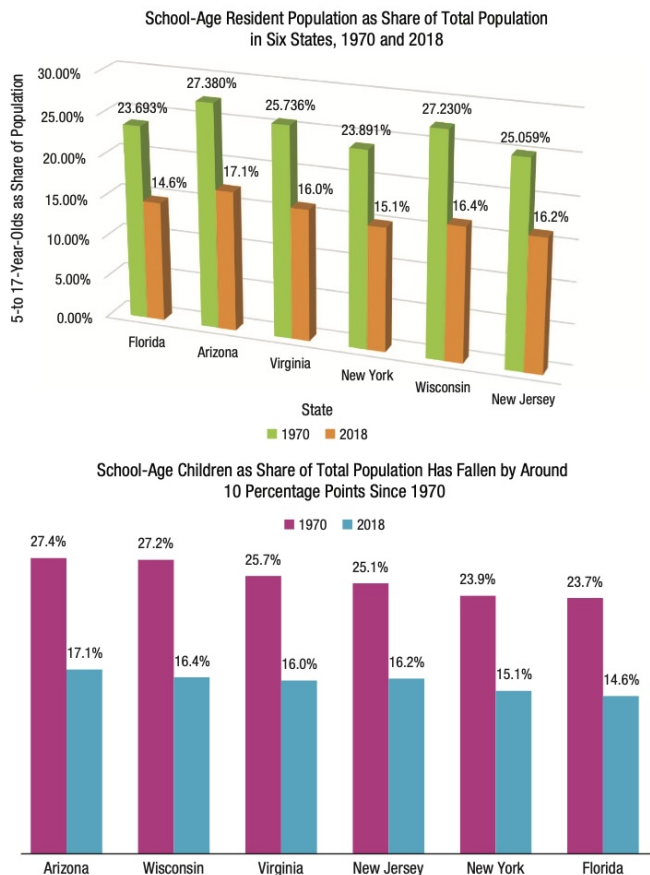
uporaba na velikih podatkih. Pri zelo velikih omrežjih postanejo povezave nepregledne in vizualizacija se spremeni v tako imenovani klobčič povezav (angl. *hairball*), kjer posameznih struktur ni več mogoče razločiti. Zato pri večjih omrežjih pogosto uporabljamo filtriranje, združevanje vozlišč ali alternativne matrične predstavitve omrežij.

Pri sodobni podatkovni analizi pogosto uporabljamo tudi kompleksnejše vizualizacijske pristope. Njihov cilj ni zgolj estetska predstavitev podatkov, ampak učinkovito razkrivanje struktur, odnosov in vzorcev, ki jih z osnovnimi grafi težko opazimo. Pri njihovi uporabi pa moramo biti posebej pozorni na ravnotežje med informativnostjo in razumljivostjo. Če v isti prikaz vključimo preveč informacij, lahko vizualizacija hitro postane nepregledna. Zato je pri načrtovanju kompleksnih vizualizacij še posebej pomembno razumevanje podatkov, naloge analize in omejitve človekovega zaznavanja.

Načela učinkovitega vizualnega oblikovanja

Dobra vizualizacija mora opazovalcu omogočiti hitro, pravilno in čim manj naporno razumevanje podatkov. Zato pri načrtovanju grafov ni pomembna le izbira tipa vizualizacije, ampak tudi način uporabe barv, organizacija informacij, poudarjanje ključnih podatkov in odstranjevanje nepomembnih elementov. Raziskave na področju vizualne percepcije kažejo, da lahko že majhne oblikovne odločitve močno vplivajo na interpretacijo podatkov. Vizualizacija mora zato podpirati človekove zaznavne sposobnosti in zmanjševati kognitivno obremenitev opazovalca. Cilj ni ustvarjanje vizualno spektakularnih grafov, ampak jasna in učinkovita komunikacija podatkov.

Nekateri primeri v tem razdelku so povzeti po članku Jonathana Schwabisha *The Practice of Visual Data Communication: What Works* iz revije *Psychological Science in the Public Interest* (2021), ki na praktičnih primerih prikazuje vpliv oblikovnih odločitev na učinkovitost vizualizacij.

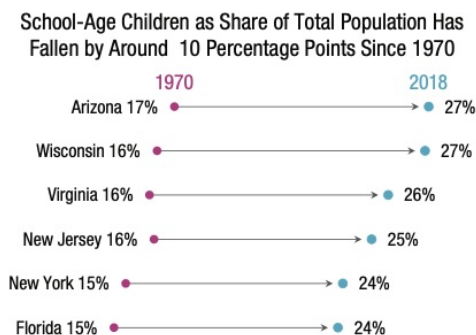


Slika 50: Primer izboljšanja stolpčnega diagrama. Zgornji graf uporablja tridimenzionalni prikaz, močne mreže in nepregledne oznake, spodnji pa enostavnejšo postavitev, neposredno označevanje podatkov in bolj pregledno uporabo barv. Povzeto po Schwabish (2021). Odstranitev nepotrebnih grafičnih elementov bistveno izboljša preglednost in usmeri pozornost na podatke namesto na dekoracijo. Takšni odvečni elementi so pogosto označeni z izrazom *chartjunk*, ki opisuje grafične podatke brez analitične vrednosti.

Preprostost in jasnost. Graf naj vsebuje le elemente, ki prispevajo k razumevanju podatkov. Nepotrebni dekorativni učinki, tridimenzionalni prikazi, ogromna omrežja ali pretirano število oznak pogosto zmanjšajo preglednost in otežijo interpretacijo. Glavno sporočilo vizualizacije mora biti razvidno takoj. Opazovalec ne bi smel ugibati, kaj graf prikazuje ali kateri podatki so pomembni. Ključne informacije morajo biti vizualno poudarjene, manj pomembni elementi pa umaknjeni v ozadje.

Vizualna hierarhija in usmerjanje pozornosti. Dobra vizualizacija mora opazovalca usmerjati k najpomembnejšim informacijam. Vizualna hierarhija določa, katere elemente zaznamo najprej in katere kasneje. To dosežemo z uporabo kontrasta, velikosti, debeline črt, položaja ali barve. Pogosto želimo poudariti le manjši del podatkov, medtem ko preostale informacije ostanejo v ozadju. Zato številni sodobni grafi uporabljajo nevtralne sive tone za večino elementov, ključne podatke pa poudarijo z izrazitejšo barvo. Tak pristop zmanjša vizualni šum in omogoči hitrejše razumevanje bistvenih informacij. Vizualna hierarhija je posebej pomembna pri kompleksnih grafih, kjer

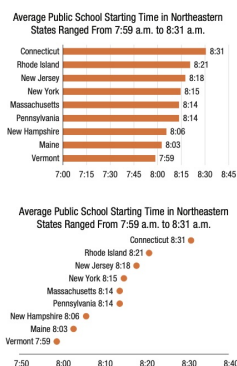
bi enakovredno poudarjanje vseh elementov povzročilo nepreglednost in povečalo kognitivno obremenitev opazovalca.



Slika 51: Točkovni prikaz uporablja nevtralne elemente in poudarjene oznake za usmerjanje pozornosti opazovalca na ključne razlike med podatki. Neposredno označevanje podatkov zmanjšuje potrebo po ločenih legendah in izboljša preglednost vizualizacije. Povzeto po Schwabish (2021).

Doslednost in zmanjševanje kognitivne obremenitve. Vizualizacije morajo uporabljati dosledne grafične konvencije. Enake barve naj predstavljajo iste kategorije, osi naj uporabljajo enake merske enote, tipografija in oznake pa naj bodo poenotene skozi celotno predstavitev.

Nedoslednosti povečajo kognitivno obremenitev, saj mora opazovalec ponovno interpretirati pomen posameznih elementov. Vizualizacija mora biti zasnovana tako, da opazovalec čim manj časa porabi za razumevanje same strukture grafa in čim več za interpretacijo podatkov. Kognitivno obremenitev zmanjšujemo tudi z neposrednim označevanjem podatkov namesto ločenih legend, z uporabo kratkih in informativnih naslovov ter z logično razporeditvijo elementov. Posebej učinkovite so vizualizacije, kjer so besedilo, oznake in grafični elementi tesno povezani.

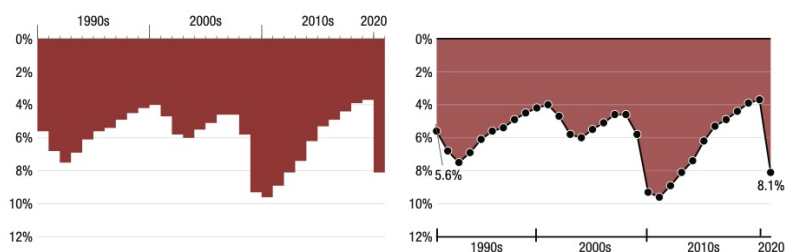


Slika 52: Primerjava stolpčnega diagrama in točkovnega prikaza. Točkovni prikaz uporablja manj vizualnega prostora in omogoča hitrejšo primerjanje vrednosti ter sprememb med kategorijami. Povzeto po Schwabish (2021).

Barve in dostopnost. Barva je eden najmočnejših vizualnih elementov, vendar jo moramo uporabljati previdno. Primerna je predvsem za razlikovanje kategorij, poudarjanje pomembnih podatkov in prikaz intenzitete vrednosti. Pomemben delež populacije ima različne oblike

motenj barvnega zaznavanja, najpogosteje težave pri razlikovanju rdeče in zelene barve. Vizualizacije morajo zato ostati razumljive tudi uporabnikom z barvno slepoto ter pri sivinskem prikazu. Zato se pogosto izogibamo problematičnim kombinacijam ter uporabljamo perceptualno primerne barvne lestvice.

Oblikovne odločitve vplivajo na interpretacijo. Na interpretacijo podatkov močno vpliva tudi način njihove vizualne predstavitve. Že majhne oblikovne odločitve lahko močno vplivajo na interpretacijo in ustvarijo napačen vtis o trendih ali razlikah med podatki. Posebej problematične so neustrezne osi, pretirano poudarjeni elementi ali neprimerna uporaba linij in površin. Takšne odločitve lahko povzročijo, da opazovalec zazna trend ali povezavo, ki v podatkih dejansko ne obstaja.



Slika 53: Dve vizualizaciji istih podatkov z obrnjeno navpično osjo. Leva različica uporablja stolpce in pravilno poudarja dolžino kot nosilko informacije, desna pa zaradi linije in položaja točk ustvarja zavajajoč vtis trenda. Povzeto po Schwabish (2021).

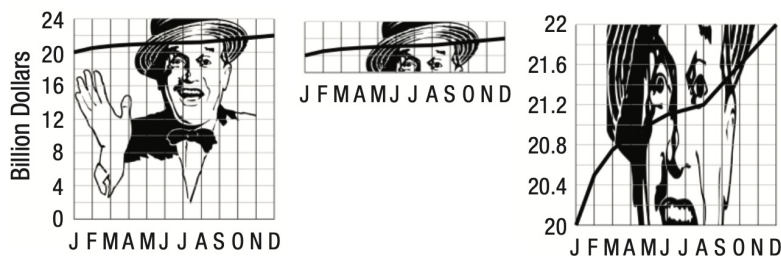
Učinkovita vizualizacija zato zahteva več kot zgolj tehnično znanje izdelave grafov. Zahteva razumevanje človekove percepcije, načel grafične komunikacije ter prilagoditev ciljnemu občinstvu. Dobra vizualizacija opazovalcu olajša razumevanje podatkov in odnosov med njimi.

Zavajajoče vizualizacije

Vizualizacije podatkov imajo velik vpliv na razumevanje informacij, zato lahko že majhne oblikovne odločitve bistveno spremenijo interpretacijo podatkov. Grafi pogosto delujejo objektivno in znanstveno, vendar lahko z neustreznim oblikovanjem ustvarijo zavajajoč vtis o velikosti razlik, trendih ali povezavah med spremenljivkami. Zaradi tega je etično oblikovanje vizualizacij pomemben del odgovorne komunikacije podatkov.

Prirezane osi in popačene skale. Eden najpogostejših načinov zavajanja je uporaba prireznih osi. Če navpična os ne začne pri nič, lahko že majhne razlike med vrednostmi delujejo bistveno večje, kot so v resnici. Takšne manipulacije pogosto pretirano poudarijo trende ali razlike med skupinami. Podoben učinek imajo popačene skale, kjer razmiki med vrednostmi niso linearni ali pa so vizualno

Nekateri primeri v tem razdelku so povzeti po članku *The Science of Visual Data Communication: What Works* (Franconeri idr., 2021), ki vključuje tudi primere napak in zavajajoče načine prikaza podatkov.



Slika 54: Primer vpliva prirezanih osi in raztegnjenih skal na interpretacijo podatkov. Enaki podatki lahko zaradi drugačne izbire osi ustvarijo bistveno drugačen vtis o velikosti razlik ali trendov. Povzeto po Franconeri in sod. (2021).

predstavljeni nesorazmerno glede na dejanske podatke. Opazovalec običajno zazna predvsem vizualno velikost razlik in manj pogosto natančno preverja numerične vrednosti na osi. Takšni prijemi so pogosti v medijih, politiki in poslovnih predstavitvah, kjer želimo določene razlike poudariti bolj, kot to upravičujejo podatki.

Zavajajoče barvne lestvice. Barva močno vpliva na zaznavanje intenzitete podatkov. Neprimerne barvne lestvice lahko ustvarijo umetne meje med vrednostmi ali poudarijo razlike, ki v podatkih niso pomembne. Posebej problematične so lestvice z močnimi prehodi med različnimi barvnimi kategorijami, na primer med modro, zeleno in rdečo. Človek zazna prehod med barvnimi kategorijami kot večjo razliko, kot dejansko obstaja v podatkih. Zaradi tega lahko kontinuirani podatki delujejo bolj diskretno ali dramatično. Pri oblikovanju vizualizacij zato uporabljamo perceptualno enakomerne barvne lestvice, kjer sprememba barve čim bolj ustreza dejanski spremembi podatkovnih vrednosti.

Prekrivanje podatkov in izbira prikazanih primerov. Pri velikih množicah podatkov lahko pride do prekrivanja točk (*overplotting*), kjer posamezne vrednosti zakrijejo druge. Posledično lahko pomembni vzorci ostanejo skriti ali pa vizualizacija daje napačen vtis gostote podatkov. Zavajajoč učinek lahko povzroči tudi selektivna izbira podatkov (angl. *cherry-picking*), kjer avtor prikaže le tiste podatke, ki podpirajo želeni zaključek. Vizualizacija lahko tako deluje prepričljivo, čeprav ne predstavlja celotne slike. Zato morajo vizualizacije jasno prikazati obseg podatkov, uporabljene filtre in morebitne omejitve pri izbiri vzorca.

Korelacija ni vzročnost. Vizualizacije pogosto učinkovito pokažejo povezave med spremenljivkami, vendar povezava sama po sebi še ne pomeni vzročne zveze. Dve spremenljivki sta lahko močno povezani, čeprav med njima ni neposrednega vpliva. Razsevni in linijski diagrami lahko hitro ustvarijo vtis vzročne povezave, posebej kadar sta podatka prikazana skupaj skozi čas. Opazovalci pogosto intuitivno sklepajo, da ena spremenljivka povzroča drugo, čeprav je povezava

lahko posledica tretjega dejavnika ali naključja. Slika 55 prikazuje, kako hitro opazimo vzorce in strukture v podatkih, tudi kadar ti niso nujno analitično pomembni. Pri interpretaciji vizualizacij moramo zato ločiti med opisom podatkov in dejanskimi vzročnimi razlagami.

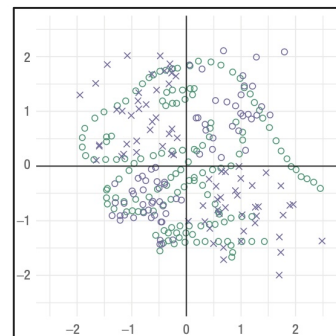
Prikazovanje negotovosti. Podatki pogosto vsebujejo negotovost, ki jo moramo ustrezno prikazati. Če negotovosti ne pokažemo, lahko vizualizacija daje lažen vtis natančnosti in gotovosti. Negotovost običajno prikazujemo z intervali zaupanja, območji verjetnosti, razponi napak ali več možnimi scenariji. Posebej pomembno je to pri vremenskih napovedih, epidemioloških modelih, ekonomskih napovedih in znanstvenih rezultatih. Vizualizacija mora zato jasno pokazati:

- katere podatke prikazujemo,
- kakšna je stopnja negotovosti,
- katere omejitve imajo podatki,
- in katerih zaključkov iz vizualizacije ne moremo zanesljivo sklepati.

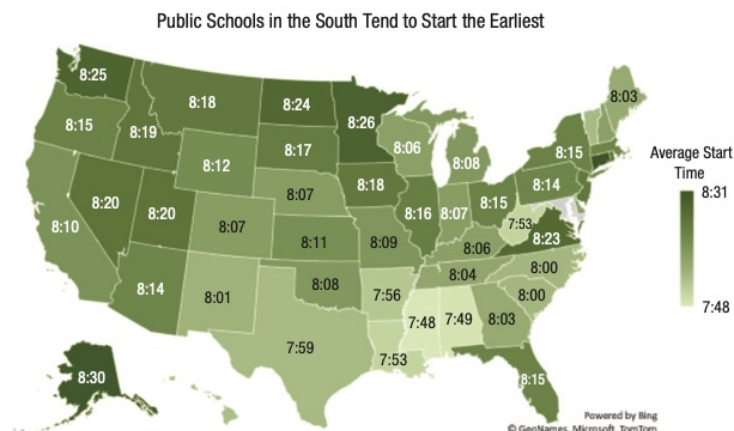
Primeri izbire in izboljšanja vizualnih predstavitev

Iste podatke lahko prikažemo na več načinov, pri čemer vsak način poudari nekoliko drug vidik: prostorski vzorec, primerjavo med enotami, povezavo med spremenljivkami, natančne vrednosti ali bolj intuitivno razumevanje časovnih podatkov. V nadaljevanju si ogledamo nekaj primerov iz članka Schwabisha (2021) v reviji *Psychological Science in the Public Interest*, ki pokažejo, kako lahko že razmeroma preproste spremembe izboljšajo razumljivost in uporabnost vizualizacije. Avtor pri tem uporablja podatke o začetku pouka v javnih šolah v ZDA ter pokaže, da lahko tudi z običajnimi orodji pripravimo precej različne, a učinkovite predstavitve podatkov.

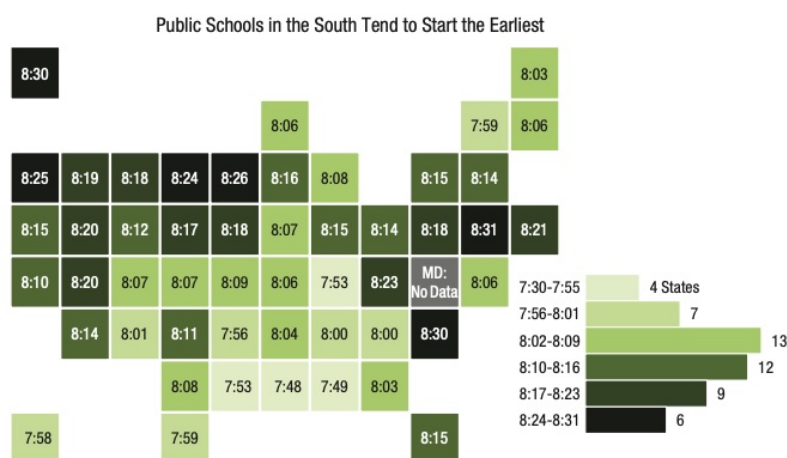
Prostorski podatki. Pri prostorskih podatkih pogosto najprej pomislimo na običajen zemljevid, vendar ta ni vedno najboljša izbira. Klasični koropletni zemljevid dobro ohranja geografsko obliko držav in zato hitro razkrije prostorske vzorce, na primer da se pouk v južnih zveznih državah začne nekoliko prej. Po drugi strani velikost geografskih območij vpliva na zaznano pomembnost podatkov: velike države na zemljevidu zavzamejo več prostora, čeprav nimajo nujno večje analitične teže. Mrežni zemljevid s ploščicami ta problem zmanjša, saj vsaki državi nameni enako velik prostor, hkrati pa omogoča dodajanje oznak ali majhnih grafičnih elementov znotraj vsake celice (slika 56). Tak prikaz žrtvuje nekaj geografske



Slika 55: Človekov vizualni sistem hitro zaznava vzorce in povezave med podatki, vendar lahko to vodi tudi do napačnega sklepanja o odnosih med spremenljivkami. Povzeto po Franconeri idr. (2021).



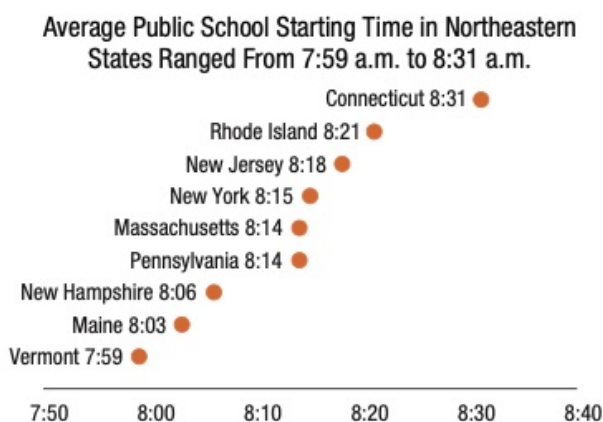
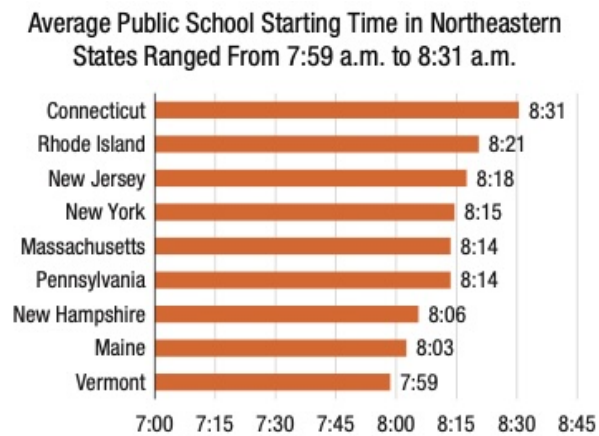
Slika 56: Dva načina prikaza prostorskih podatkov o povprečnem začetku pouka v javnih šolah v ZDA: običajni koropletni zemljevid in mrežni zemljevid s ploščicami. Prvi bolje ohranja geografsko obliko, drugi pa vsaki državi nameni enak prostor in tako olajša primerjavo med državami.



natančnosti, vendar lahko izboljša primerljivost in poveča uporabnost prikaza za bralca.

Stolpci ali točke. Stolpčni diagram je ena najbolj znanih in razumljivih vizualizacij za primerjanje vrednosti med skupinami. Njegova prednost je, da dolžine stolpcev primerjamo razmeroma natančno, še posebej, kadar imajo skupno izhodišče. Toda stolpčni diagrami lahko postanejo vizualno težki, posebej pri večjem številu kategorij ali daljših oznakah. Točkovni diagram za iste podatke pogosto uporablja manj grafičnega prostora in pusti več prostora za oznake, komentarje ali neposredno označevanje vrednosti (slika 57). Primer zato ne kaže, da je stolpčni diagram napačen, ampak da lahko preprostejša in lažja grafična oblika bolj berljiva, kadar nas zanima predvsem vrstni red in primerjava posameznih vrednosti.

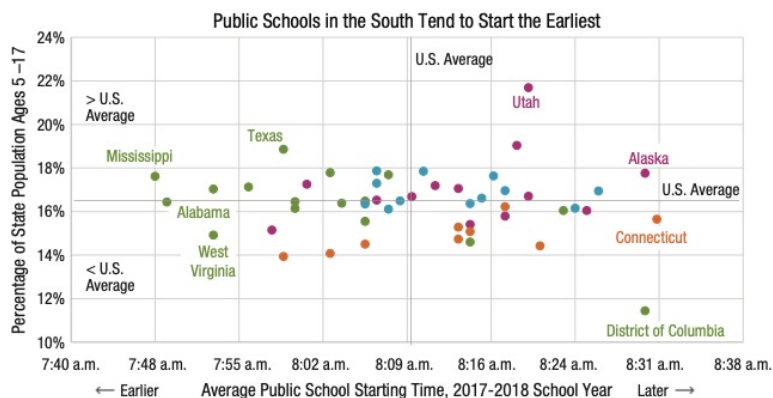
Razsevni diagram kot prostor za razlago. Razsevni diagram je osnovni prikaz za raziskovanje povezave med dvema numeričnima



Slika 57: Primerjava stolpčnega diagrama in točkovnega prikaza za iste podatke. Stolpci omogočajo neposredno primerjavo dolžin, točkovni prikaz pa zmanjša vizualno težo grafa in omogoči preglednejše označevanje vrednosti.

spremenljivkama. Vendar dober razsevni diagram ni le oblak točk. Z dodatnimi oznakami, povprečnimi črtami, pojasnili osi in označenimi osamelci lahko bralcu pomagamo razumeti, kaj naj v grafu opazi. Pri podatkih o začetku pouka in deležu šoloobveznih otrok v populaciji lahko vodoravne in navpične referenčne črte razdelijo prostor grafa na območja nad in pod povprečjem, oznake izbranih držav pa usmerijo pozornost na primere, ki posebej odstopajo (slika 58). Pri takem prikazu je pomembno ravnotežje: preveč oznak graf obremeni, premišljene oznake pa lahko bistveno izboljšajo interpretacijo.

Tudi tabela je vizualizacija. Tabele pogosto razumemo kot nasprotje grafov, vendar so tudi tabele oblika vizualne predstavitve podatkov. Slabo oblikovana tabela oteži primerjanje vrednosti: glave stolpcev niso jasno ločene, besedilo in številke niso ustrezno poravnani, število decimalnih mest je nedosledno, enote pa so zapisane neenotno. Z razmeroma majhnimi popravki lahko tabela postane bistveno preglednejša: glave stolpcev so jasno ločene, besedilo je levo poravnano,



Slika 58: Razsevni diagram povezave med povprečnim začetkom pouka in deležem prebivalstva v starosti od 5 do 17 let. Referenčne črte, oznake osi in izbrane oznake držav pomagajo bralcu razumeti strukturo podatkov ter prepoznati osamelce.

State	Total population		Students		Percent		Average start time	% Before 8 a.m.
	1970	2018	1970	2018	1970	2018		
Arizona	1,775.00	7048.876	486.0	1202.589	27.38%	17%	8:14	29.5%
Florida	6,791.00	20976.812	1609.0	3063.834	23.69%	14.61%	8:15	26.4%
New Jersey	7,171.00	8888.543	1797.0	1442.509	25.06%	16.23%	8:18	23.5%
New York	18,241.00	19590.719	4358.0	2954.768	23.89%	15%	8:15	27%
Virginia	4,651.00	8465.207	1197.0	1358.561	25.74%	16.05%	8:23	17.8%
Wisconsin	4,418.00	5792.051	1203.0	947.144	27.23%	16.35%	8:06	28.9%

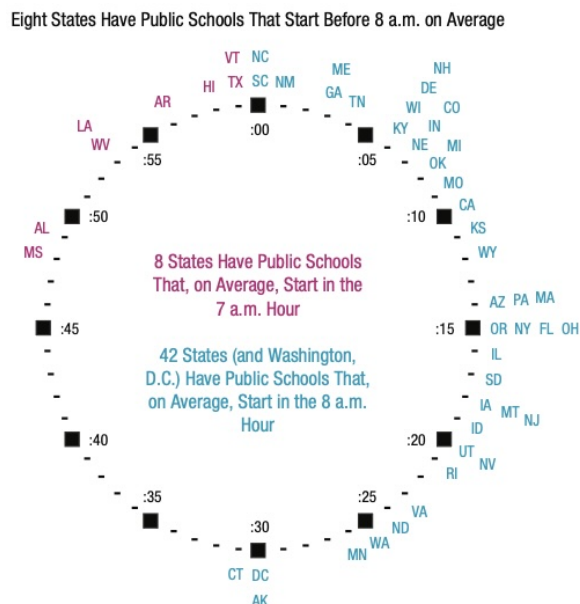
	Total Population (thousands)		Students (thousands)		Students as share of total population (%)		Average start time	Percentage starting before 8 a.m.
	1970	2018	1970	2018	1970	2018		
Arizona	1,775	7,049	486	1,203	27.4	17.0	8:14	29.5
Florida	6,791	20,977	1,609	3,064	23.7	14.6	8:15	26.4
New Jersey	7,171	8,889	1,797	1,443	25.1	16.2	8:18	23.5
New York	18,241	19,591	4,358	2,955	23.9	15.0	8:15	27.0
Virginia	4,651	8,465	1,197	1,359	25.7	16.0	8:23	17.8
Wisconsin	4,418	5,792	1,203	947	27.2	16.4	8:06	28.9

Slika 59: Primer izboljšanja tabele. Preglednejša različica uporablja jasnejšo hierarhijo glave, ustrezno poravnavo besedila in števil, enotno število decimalnih mest ter majhne stolpčne prikaze za hitrejšo zaznavanje vzorcev. Povzeto po Schwabish (2021), Figure 10.

številke desno poravnane, natančnost zapisa je poenotena, dodani majhni stolpci v zadnjem stolpcu pa pomagajo hitro zaznati vzorec (slika 59). Tak primer je posebej pomemben za poročila in znanstvena besedila, kjer tabele pogosto nosijo veliko informacij.

Vizualizacija, ki izhaja iz pomena podatkov. Kadar podatki opisujejo čas dneva, lahko vizualizacija izkoristi obliko, ki je bralcu že znana iz vsakdanjega življenja. Namesto zemljevida, stolpčnega grafa ali tabele lahko začetke pouka razporedimo po številčnici ure, kjer položaj oznake neposredno ustreza času začetka (slika 60). Tak prikaz ni nujno najbolj standarden, vendar je dobro usklajen s pomenom podatkov: bralec takoj razume, da gre za čas, in lahko poišče posamezno državo na znani krožni strukturi. Primer kaže, da je lahko nekoliko bolj igriva ali nestandardna vizualizacija še vedno analitično smiselna, če podpira razumevanje podatkov in ne zakriva njihovega pomena.

Skupno sporočilo teh primerov je, da učinkovita vizualizacija nastane iz povezave med podatki, nalogo in občinstvom. Ni dovolj, da izberemo prvi graf, ki ga ponudi programsko orodje. Premisliti



Slika 60: Alternativni prikaz časovnih podatkov s številčnico ure. Oznake držav so razporejene glede na povprečni čas začetka pouka, barva pa dodatno loči države, kjer se pouk v povprečju začne pred osmo uro.

moramo, kaj želimo poudariti, katere primerjave naj bralec opravi, koliko razlage potrebuje in kateri prikaz bo podatke predstavil najbolj jasno. Včasih je to običajen stolpčni diagram, drugič zemljevid, tabela, točkovni prikaz ali celo ura.

Vizualna analitika in interaktivnost

Sodobna vizualizacija podatkov ni več omejena le na statične slike v tiskanih medijih (članki, poročila) ali na drsnicah. Vedno pomembnejšo vlogo imajo interaktivne vizualizacije, pri katerih lahko uporabnik podatke raziskuje, spreminja pogled na podatke in se osredotoča na zanimive podmnožice podatkov. Tak pristop uporabljajo tehnike vizualne analitike, ki združujejo vizualizacijo, interakcijo in analitično raziskovanje podatkov.

Ko uporablja vizualne vmesnike, uporabnik ni le pasivni opazovalec grafične predstavitve, ampak lahko aktivno spreminja prikaz podatkov in, kolikor mnu vmesnik dopušča, preiskuje in raziskuje podatke. Najpogostejše interakcije, ki jih taki sistemi implementirajo, vključujejo:

- **filtriranje**, kjer izberemo le del podatkov,
- **brushing**, kjer označimo podmnožico podatkov v enem prikazu,
- **povečevanje in premikanje** (*zooming in panning*), kjer se lahko osredotočimo na del podatkov,

- **izbiro elementov** (*selection*), kjer v eni vizualizaciji izberemo elemente, ki jih potem navadno prikažemo na dodatni vizualizaciji,
- **podrobnosti na zahtevo** (*details on demand*), kjer se dodatne informacije prikažejo ob kliku miške ali ko stojimo z miško na določenem elementu, za katerega nas zanimajo dodatne podrobnosti, in
- **barvanje in označevanje skupin**, kjer na grafični način uporabnik razdeli podatke v njemu zanimive skupine, za katere je potem seveda potrebno prikazati razlike z dodatnimi prikazi ali v dodatnih vizualizacijah.

V preiskovalne namene so posebej uporabne povezane vizualizacije (*linked visualizations*). Pri takem pristopu izbira podatkov v enem grafu (avtomatično in hipno) vpliva tudi na druge prikaze. Če na primer v razsevnem diagramu označimo določeno skupino podatkov, se isti podatki istočasno označijo tudi v histogramu ali tabeli. Tak način omogoča učinkovito raziskovanje večdimenzionalnih podatkov in hitro odkrivanje povezav med atributi.

Interaktivne vizualizacije pogosto združujemo v nadzorne plošče oziroma *dashboards*. Ti prikazi vsebujejo več med seboj povezanih grafov, filtrov in tabel, ki uporabniku omogočajo pregled nad podatki in interaktivno analizo. Podoben pristop uporabljajo tudi sistemi za vizualno podatkovno analitiko, kot je ljubljanski Orange Data Mining, kjer uporabnik gradi analitični potek dela (*workflow*) s povezovanjem posameznih komponent za uvoz podatkov, analizo, modeliranje in vizualizacijo in na ta način določi način obdelave podatkov ter med sabo poveže vizualizacije.

Podpora za interaktivno vizualizacijo danes ponujajo številna programska orodja. Tableau je eno najbolj razširjenih okolij za pripravo interaktivnih nadzornih plošč in raziskovalno analizo podatkov, saj omogoča hitro povezovanje grafov, filtrov in interaktivnih pogledov brez programiranja. Podobno vlogo ima Microsoft Power BI, ki je močno povezan z ekosistemom podatkovnih storitev podjetja Microsoft in se pogosto uporablja za poslovno analitiko ter spremljanje podatkov v realnem času. Orodja, kot je Orange Data Mining, KNIME in RapidMiner, pa so bolj usmerjena v raziskovalno analizo in strojno učenje, kjer uporabnik z vizualnim povezovanjem komponent gradi celoten analitični potek dela. Skupna značilnost teh sistemov je podpora interaktivnosti, povezanim vizualizacijam ter postopnemu raziskovanju podatkov.

Poseben izziv predstavljajo podatki v realnem času oziroma pretocni podatki (*streaming data*). Pri takih podatkih se vizualizacija sproti posodablja, ko prihajajo novi podatki. Takšne prikaze srečamo pri spremljanju omrežnega prometa, finančnih trgov, senzorjev ali

družbenih omrežij. Poleg samega prikaza podatkov je pri takih sistemih pomembna tudi hitrost osveževanja, poudarjanje pomembnih sprememb ter preprečevanje preobremenitve uporabnika z informacijami.

Interaktivnost zato ni le estetski dodatek, ampak pomembno orodje za raziskovanje podatkov. Dobro zasnovane interaktivne vizualizacije uporabniku omogočajo, da sam raziskuje podatke, postavlja vprašanja in postopoma gradi razumevanje analiziranega problema.

Pripovedovanje zgodb s podatki

Vizualizacija podatkov je predvsem način komunikacije o spoznanjih, ki smo jih na osnovi podatkov pridobili. Pri pripravi vizualizacije zato običajno pričnemo z vprašanjem, kaj želimo sporočiti. Isti podatki lahko podprejo različne zgodbe: lahko poudarimo trend, primerjavo, izjemen primer, negotovost ali spremembo skozi čas. Naloga avtorja vizualizacije je, da izbere prikaze ali redosled prikazov, ki podprejo glavno sporočilo, ter usmeri pozornost bralca na pomembne dele podatkov in vzorce. Pri tem imajo pomembno vlogo barva, kontrast, velikost elementov, označevanje in komentarji (*annotation*), s katerimi lahko poudarimo ključne informacije in zmanjšamo vpliv manj pomembnih podrobnosti.

Učinkovita vizualizacija podatkov zato pogosto deluje kot pripoved (*storytelling with data*). Bralca vodi skozi podatke, postopoma razkriva pomembne vzorce in pomaga pri interpretaciji rezultatov. Tak pristop je posebej značilen za novinarske in spletne interaktivne vizualizacije, kjer uporabnik podatke raziskuje korak za korakom. Pri tem moramo vedno upoštevati tudi občinstvo: vizualizacije za strokovnjake lahko vsebujejo več podrobnosti in kompleksnejše prikaze, medtem ko morajo biti vizualizacije za širšo javnost bolj neposredne in hitro razumljive. Dobra vizualizacija zato ne prikazuje le podatkov, ampak pomaga oblikovati razumevanje problema.

Programska orodja za gradnjo vizualizacij

Vizualizacije podatkov danes pogosto ne nastajajo več kot ročno narisane slike, temveč kot rezultat programskega opisa podatkov, njihovih preslikav v vizualne elemente in pravil interakcije. Tak pristop je posebej pomemben v podatkovni znanosti, kjer želimo grafe graditi ponovljivo, jih prilagajati novim podatkom in jih vključevati v poročila, spletne strani ali interaktivna analitična okolja.

Najpreprostejši primer takega pristopa v Pythonu je uporaba knjižnice `matplotlib`. Ta omogoča natančen nadzor nad posameznimi

elementi grafa in je osnova številnih drugih vizualizacijskih knjižnic.

```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [2.1, 2.8, 3.6, 3.9, 5.1]

plt.figure()
plt.scatter(x, y)
plt.xlabel("število ur učenja")
plt.ylabel("rezultat")
plt.savefig("scatter.pdf", bbox_inches="tight")
```

Koda zgradi enostaven razsewni diagram, kjer vsaka točka predstavlja en primer. Tak prikaz je tipičen za raziskovanje povezave med dvema numeričnima spremenljivkama. Knjižnica `matplotlib` je nekoliko bolj nizkonivojska, zato moramo sami določiti osi, oznake, legende in druge grafične elemente.

Za statistične prikaze pogosto uporabimo knjižnico `seaborn`, ki gradi na `matplotlibu`, vendar ponuja višjenivojske funkcije za pogoste tipe grafov.

```
import seaborn as sns
import pandas as pd

d = pd.DataFrame({
    "skupina": ["A", "A", "A", "B", "B", "B"],
    "vrednost": [4.1, 5.0, 4.7, 6.2, 5.8, 6.5],
})

sns.boxplot(data=d, x="skupina", y="vrednost")
plt.savefig("boxplot.pdf", bbox_inches="tight")
```

V tem primeru zgradimo škatlo z brki za primerjavo porazdelitev med dvema skupinama. Prednost knjižnice `seaborn` je, da neposredno razume podatkovne tabele, kakršne uporabljamo v knjižnici `pandas`, in zato omogoča hitro gradnjo raziskovalnih grafov.

Drugačen pristop uporablja knjižnica `Altair`. Ta temelji na jeziku `Vega-Lite`, kjer vizualizacijo opišemo deklarativno: določimo podatke, vrsto grafa in preslikave stolpcev v vizualne elemente, knjižnica pa iz tega sestavi končno vizualizacijo.

```
import altair as alt
import pandas as pd

d = pd.DataFrame({
    "ure": [1, 2, 3, 4, 5],
    "rezultat": [2.1, 2.8, 3.6, 3.9, 5.1],
```

```

})

chart = alt.Chart(d).mark_point().encode(
    x="ure",
    y="rezultat",
)

chart.save("altair-scatter.html")

```

Pri tem programu ne določamo neposredno, kako naj se izriše vsaka točka. Namesto tega povemo, da naj bodo vrednosti stolpca *ure* prikazane na osi *x*, vrednosti stolpca *rezultat* pa na osi *y*. Tak deklarativni pristop je bližje formalnim jezikom za opis vizualizacij.

Med najpomembnejšimi formalnimi jeziki za opise vizualizacij danes štejemo Vega in Vega-Lite. Oba uporabljata zapise v obliki JSON. Vega-Lite je višjenivojski in primeren za običajne statistične grafe, Vega pa omogoča podrobnejši nadzor nad transformacijami podatkov, interakcijami in grafičnimi elementi. Knjižnica Altair je pravzaprav Pythonov vmesnik za gradnjo Vega-Lite specifikacij.

Za bolj proste in unikatne spletne vizualizacije se pogosto uporablja D3.js. Ta ni deklarativni jezik v istem smislu kot Vega, temveč JavaScript knjižnica, ki podatke poveže z elementi spletne strani, na primer z elementi SVG. D3 omogoča zelo veliko prilagodljivost, vendar zahteva tudi več programiranja.

Interaktivne vizualizacije lahko v Pythonu gradimo tudi s knjižnico plotly. Njeni grafi so spletni objekti, zato jih lahko odpremo v brskalniku, vključimo v spletno stran ali uporabimo v interaktivnih zvezkih.

```

import plotly.express as px
import pandas as pd

d = pd.DataFrame({
    "ure": [1, 2, 3, 4, 5],
    "rezultat": [2.1, 2.8, 3.6, 3.9, 5.1],
    "skupina": ["A", "A", "B", "B", "B"],
})

fig = px.scatter(
    d, x="ure", y="rezultat", color="skupina",
    hover_data=["skupina"]
)

fig.write_html("plotly-scatter.html")

```

Interaktivne vizualizacije lahko gradimo tudi deklarativno, kjer ne opisujemo posameznih grafičnih elementov, temveč povezave med podatki, vizualnimi atributi in interakcijami. Tak pristop pose-

bej dobro podpira knjižnica Altair, ki temelji na jeziku Vega-Lite. Naslednji primer prikazuje dva povezana grafa. V levem razsevnem diagramu lahko uporabnik z miško izbere skupino točk, histogram na desni pa nato prikaže porazdelitev samo za izbrane primere.

```
import altair as alt
from vega_datasets import data

d = data.cars()

brush = alt.selection_interval()

scatter = alt.Chart(d).mark_point(size=60).encode(
    x="Horsepower:Q",
    y="Miles_per_Gallon:Q",
    color="Origin:N"
).add_params(brush)

hist = alt.Chart(d).mark_bar().encode(
    x=alt.X("Miles_per_Gallon:Q", bin=True),
    y="count()",
    color="Origin:N"
).transform_filter(brush)

chart = scatter | hist
chart.save("linked-view.html")
```

Zgoraj smo sicer vizualizacijo shranili v datoteko HTML, a še bolj primerno bi bilo kodo uporabiti v interaktivnih okoljih kot je Marimo.

Vizualizacija podatkov in umetna inteligenca

Čeprav smo se v vseh poglavjih do tega ukvarjali s strojnim učenjem in torej podlagami za umetno inteligenco, o povezavi te z vizualizacijami nismo napisali ničesar. Čas je, da v tej smeri zaključimo. Namreč, vizualizacija podatkov in umetna inteligenca skupaj odpirata izjemno širok prostor novih pristopov in raziskovalnih vprašanj. Umetna inteligenca lahko pomaga pri samodejnem oblikovanju vizualizacij, izbiri ustreznih prikazov ter prilagajanju predstavitev različnim uporabnikom in kontekstom. Vizualizacije vse bolj postajajo del pogovornih vmesnikov, kjer modeli ne odgovarjajo več le z besedilom, temveč tudi z dinamičnimi grafičnimi prikazi, ki podpirajo razlago, raziskovanje podatkov in skupno razmišljanje. Umetna inteligenca bo imela izjemno pomembno vlogo pri razlagi podatkov in razložljivih vizualizacijah, torej vizualizacijah, kjer bomo s prijemi umetne inteligence v vizualizacije vključevali vsebine, ki te dodatno

razložijo, ali pa sploh poiskali vizualizacije, ki so razložljive in odgovorijo na uporabnikovo vprašanje.

Pomembno področje postaja tudi oblikovanje zgodb z uporabo podatkov, kjer se prepletajo analiza, pisna zgodba in vizualna komunikacija. Ob tem se odpirajo še številne druge možnosti: interaktivna razlaga kompleksnih modelov, generiranje vizualnih analitičnih povzetkov, sodelovanje med človekom in modelom pri raziskovanju podatkov ter razvoj novih načinov vizualnega razmišljanja, kjer umetna inteligenca ne nastopa le kot orodje za analizo, temveč kot sogovornik in soustvarjalec interpretacij.