Vizualizacija podataka

Uvod u znanost o podacima

3. predavanje

Ak. god. 2021./2022.

Vizualizacija podataka

- Engl. *data visualization*

Dvije glavne svrhe:

- **U analizi podataka**
  - Podupire rasuđivanje o prisutnim informacijama
- **U komunikaciji**
  - Informira i uvjerava druge sudionike

Source: Yvette, W., Pixabay

Vizualizacija podataka za analizu podataka

- Ciljevi vizualizacije u svrhu analize podataka:
  - Otkrivanje odnosa među varijablama i među primjerima varijabli
  - Sažeti opis vrijednosti varijabli
  - Bolje razumijevanje rezultata analize
- Vizualizacija je vrlo bitna za **razumijevanje skupa podataka** (vrsti varijabli, veličine skupa primjera), otkrivanje pogrešaka u podacima i postavljanje ostvarivih ciljeva analize
  - Neki ciljevi nisu ostvarivi ako je skup podataka neadekvatan (premalen, prevelik, sadrži neke vrste varijabli, a neke ne i sl.)
  - Mnoge pogreške u podacima se teško otkrivaju bez vizualizacije

Vizualizacija podataka za komunikaciju

- Ciljevi vizualizacije u svrhu komunikacije:
  - Zaokuplja pozornost i uključuje sudionike mnogo bolje od teksta/brojaka
  - Omogućuje pričanje priče na vizualan način
  - Omogućuje fokus na pojedine zanimljive aspekte, sakrivajući detalje (apstrakcija)

- Vizualizacija je vrlo bitna za uključivanje drugih ljudi koji nisu radili na skupu podataka
  - Približava im problematiku skupa podataka i najvažnije rezultate
  - Olakšava raspravu i donošenje poslovnih odluka

Vizualizacija podataka

- **Statička vizualizacija**
  - Izvrsna za istraživanje podataka
  - Razvija se tijekom zadnjih par stoljeća
  - **U fokusu ovog predavanja**
- **Interaktivna vizualizacija**
  - Omogućuje korisniku **akcije za izmjenu elemenata** u grafičkom prikazu
  - Sve češća prilikom prikaza rezultata (npr. nadzorna ploče, engl. *dashboard*)
  - Temelji se na novim radnim okvirima za web
  - https://www.omnisci.com/technical-glossary/interactive-data-visualization
  - https://coronavirus.jhu.edu/map.html
  - https://towardsdatascience.com/dashboards-are-dead-b9f12eeb2ad2

Za one koji žele znati više

- Predmet na FER-u cijeli posvećen vizualizaciji podataka:

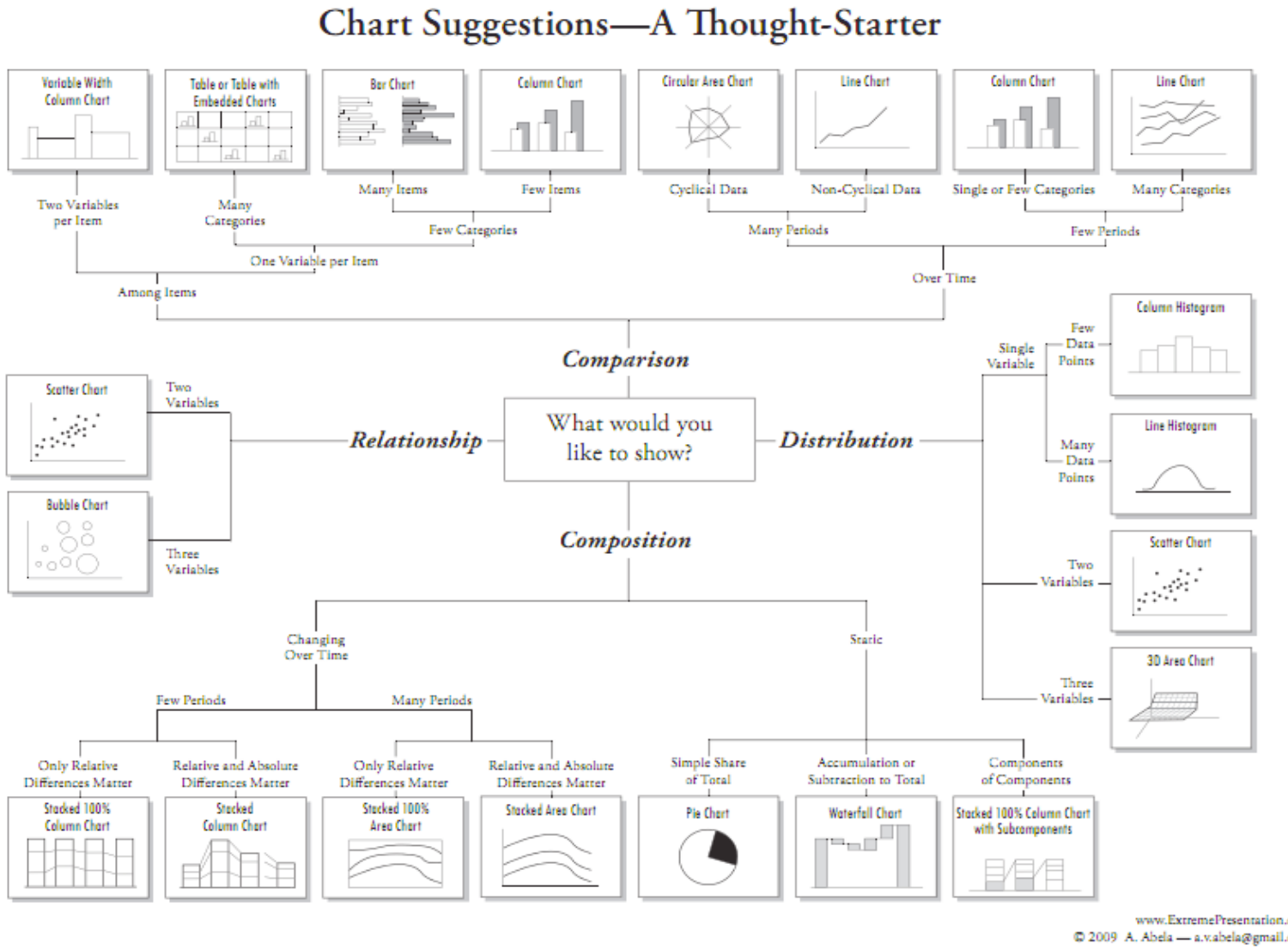
- https://www.fer.unizg.hr/predmet/vizpod

Sadržaj

- Navigacija kroz grafove za vizualizaciju podataka
- Principi i najbolje prakse vizualizacije
- Primjeri korištenja vizualizacije
- Alati za vizualizaciju

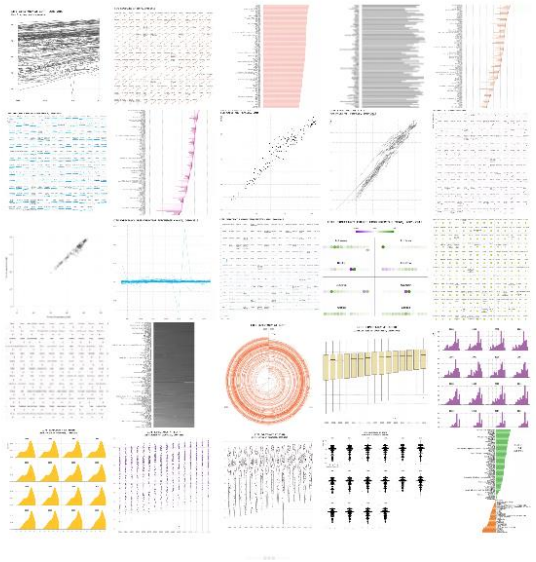
Navigacija kroz grafove za vizualizaciju podataka

Izbor grafova



- Veliki broj dostupnih grafova
- Izbor uvelike ovisi o tome:
  - Kakve podatke imamo?
  - Što točno želimo prikazati iz podataka?
- Različiti alati podupiru različiti broj i vrste grafova

Primjer ekstremne vizualizacije



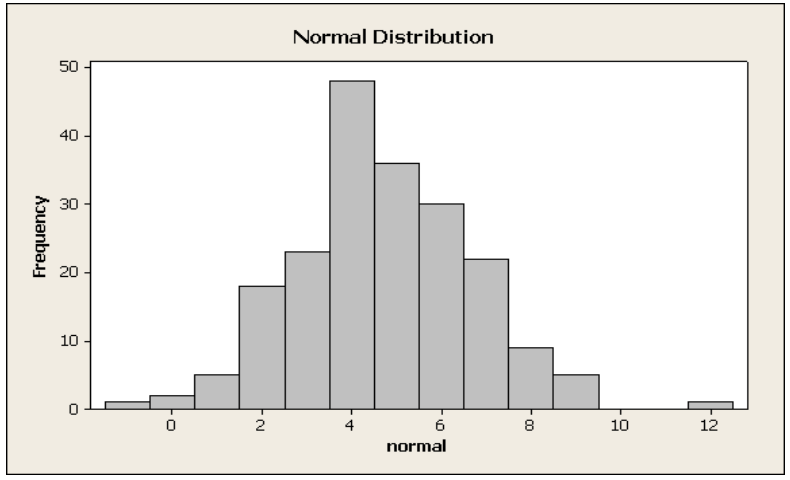
- Jedan skup podataka, sadrži očekivanu životnu dob po zemljama svijeta (WHO, 2000.-2015.), vizualiziran na **25** načina
- http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways
- Svaki graf nudi jedinstveni pogled na isti skup
- Edukativno, ali uglavnom nepotrebno u praksi
- Ispravan izbor grafa za vizualizaciju traži **iskustvo** u analizi i prikazu podataka, a najbolje je početi s osnovnim grafovima

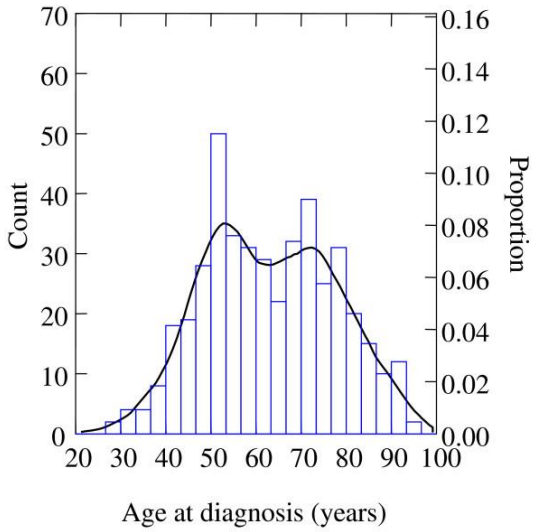
Najčešći grafovi s obzirom na broj varijabli

- **Jedna varijabla**
  - Histogram, kutijasti graf (engl. *box plot, box-and-whisker plot*), pitni graf (engl. *pie chart*), stupčasti graf (engl. *column chart*)
- **Dvije varijable**
  - Graf raspršenja (engl. *scatter plot*), linijski graf (engl. *line chart*)
- **Više od dvije varijable**
  - Matrica grafa raspršenja (engl. *scatter plot matrix*), posloženi graf (engl. *stacked plot*), mjehuričasti graf (engl. *bubble chart*), površinski 3D graf (engl. *3D area chart, surface chart*), radarski graf (engl. *radar chart, spider chart, web chart, star chart*)

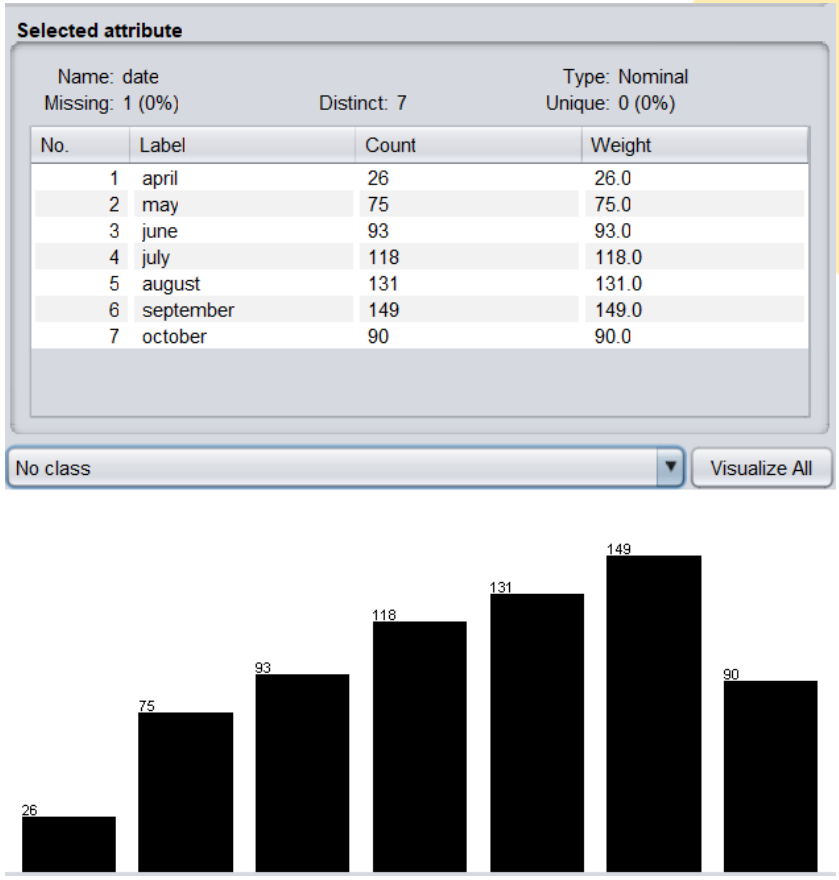
Histogram

- Prikaz pojedinačnih varijabli
- **Kategoričke** (uobičajeni histogram) i
- **Numeričke** (diskretizacija vrijednosti ili linijski histogram)



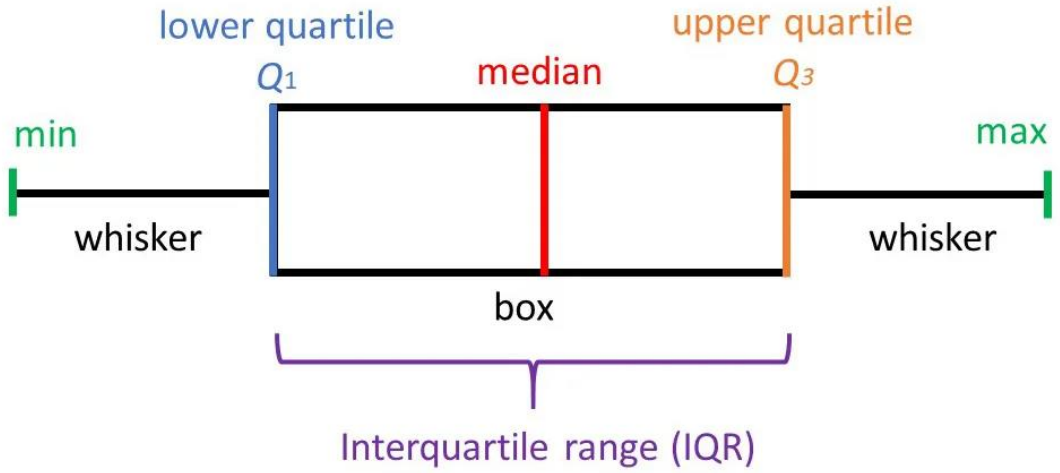


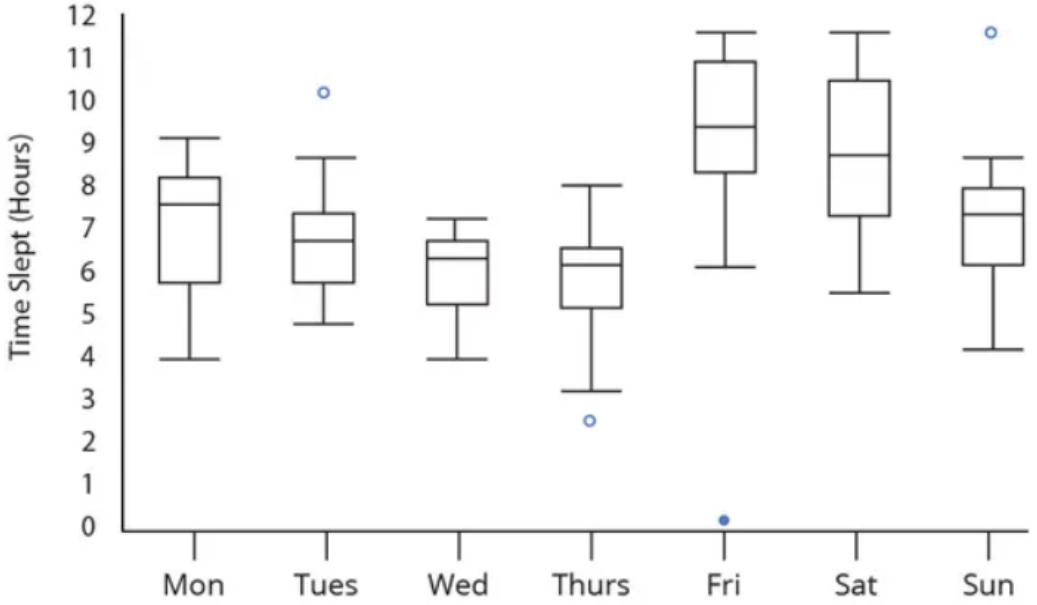
Source: ADA, 3rd lect., EPFL, 2020.



Kutijasti dijagram

- Prikaz pojedinačne **numeričke** varijable, ali moguće i više njih radi usporedbe
- Pogodno za prikaz razine ukošenosti (engl. *skeweness*) i varijabilnosti (engl. *variability*) razdiobe



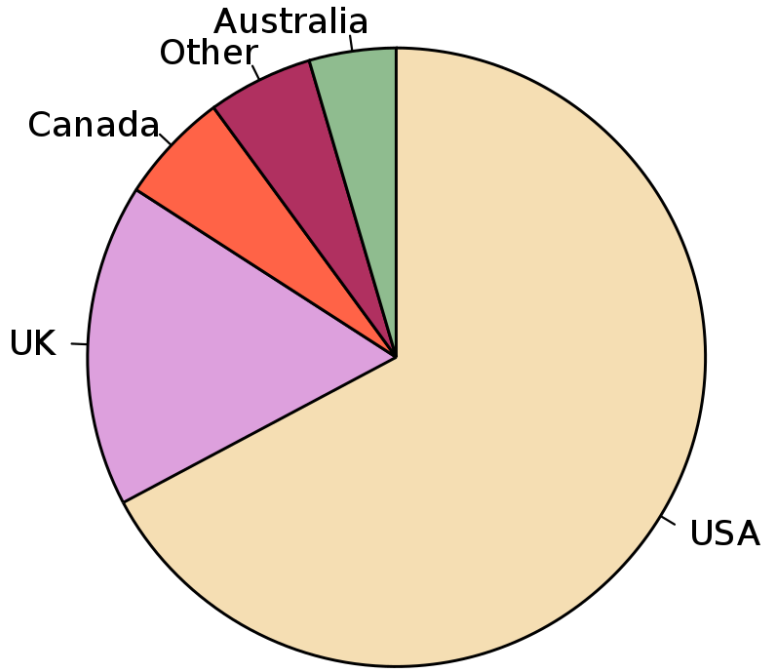


Source: McLeod, S. A. (2019, July 19). What does a box plot tell you? Simply psychology: https://www.simplypsychology.org/boxplots.html

Pitni dijagram

- Prikaz pojedinačnih varijabli
- **Kategoričke** varijable
- Prikazuje relativni odnos broja primjera pojedinih kategorija određene varijable

- U primjeru: broj ljudi kojima je materinji jezik engleski, kategorije su zemlje svijeta

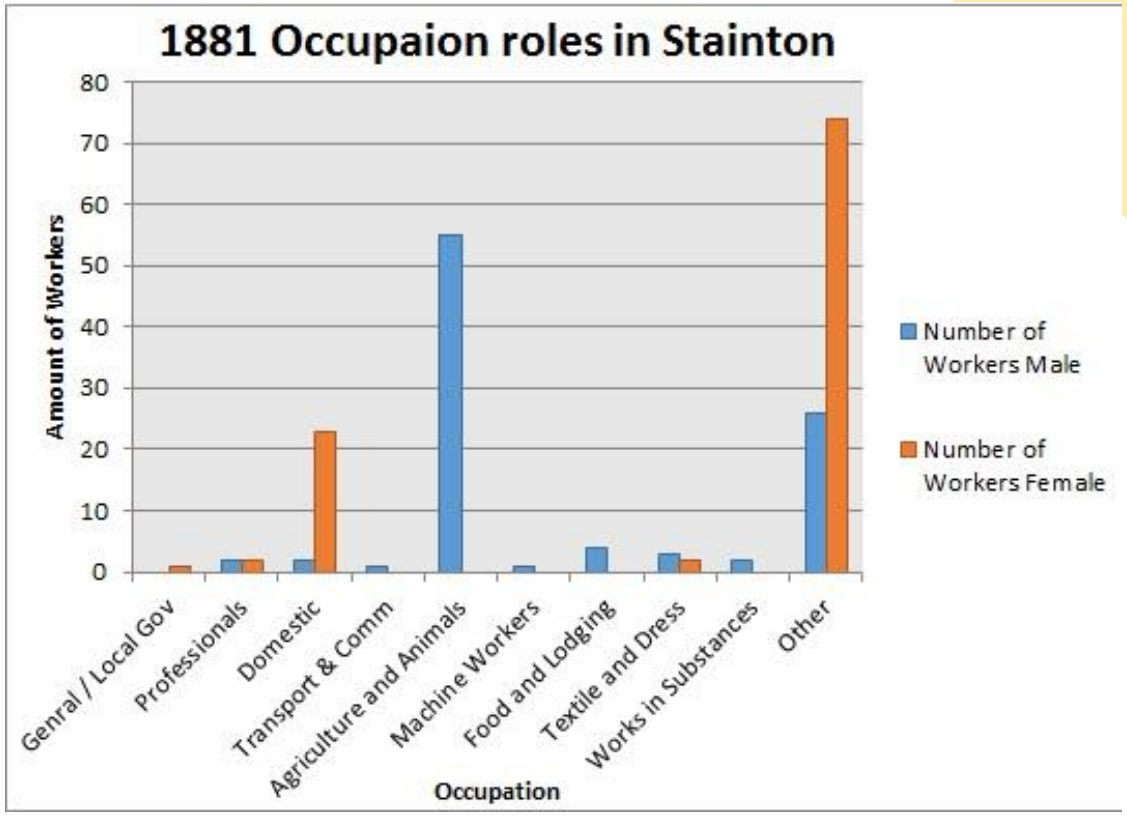


Source: https://en.wikipedia.org/wiki/Pie_chart#/media/File:English_dialects1997.svg

Stupčasti dijagram

- Prikaz pojedinačnih **kategoričkih** varijabli
- Svaka kategorija ima svoj naziv i jedan ili više pridruženih stupaca
- Stupaca ima više od jedan ako se paralelno razmatra utjecaj neke druge kategoričke varijable na onu koja se prikazuje

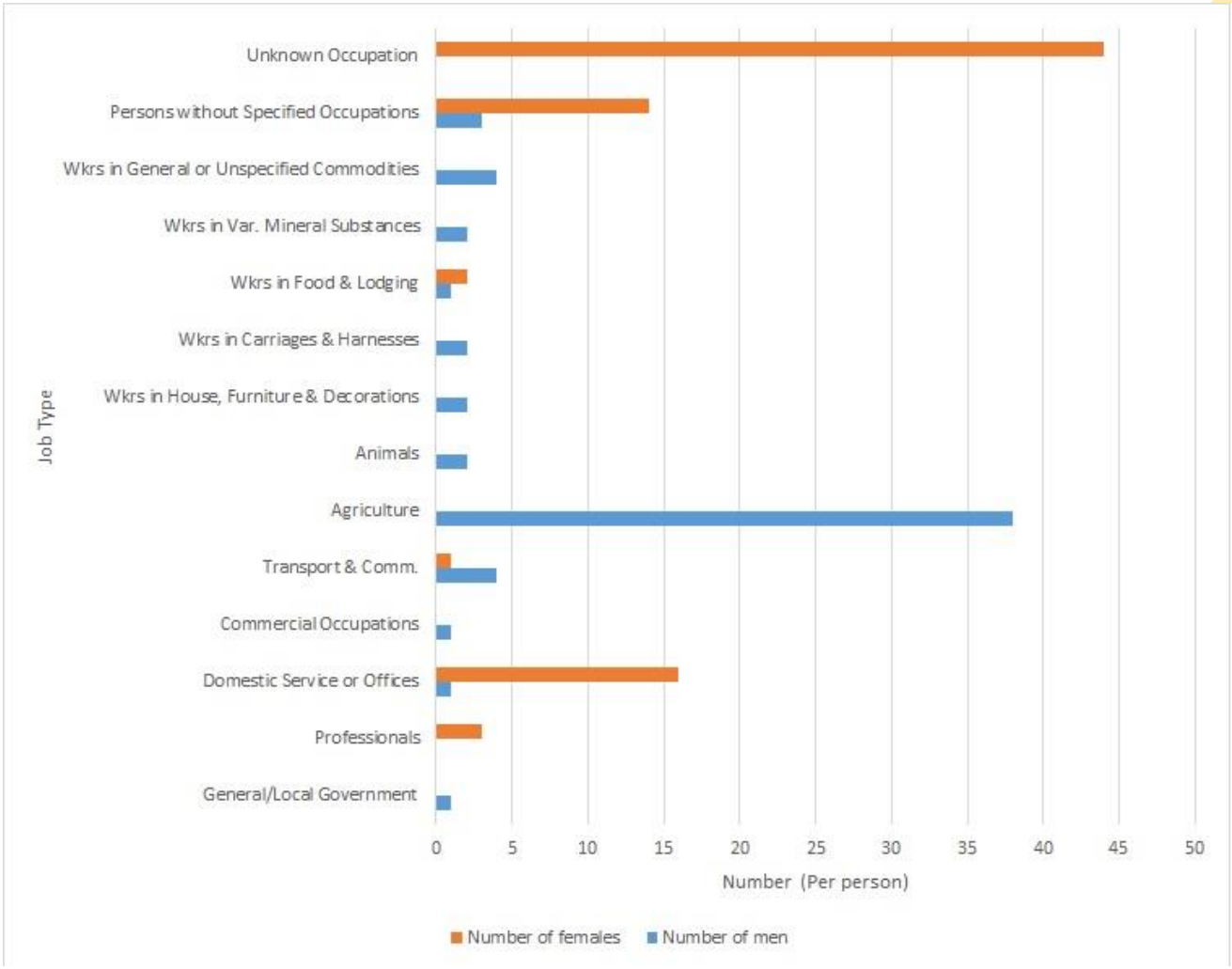
- U primjeru: varijabla zanimanje (Occupation) ima 10 kategorija, svaka kategorija prikazuje broj primjera za muške i za ženske radnike (utjecaj varijable spol)



Source: https://commons.wikimedia.org/wiki/Category:Demographic_bar_charts#/media/File:1881_bar_chart_paint.jpg

Horizontalni stupčasti dijagram

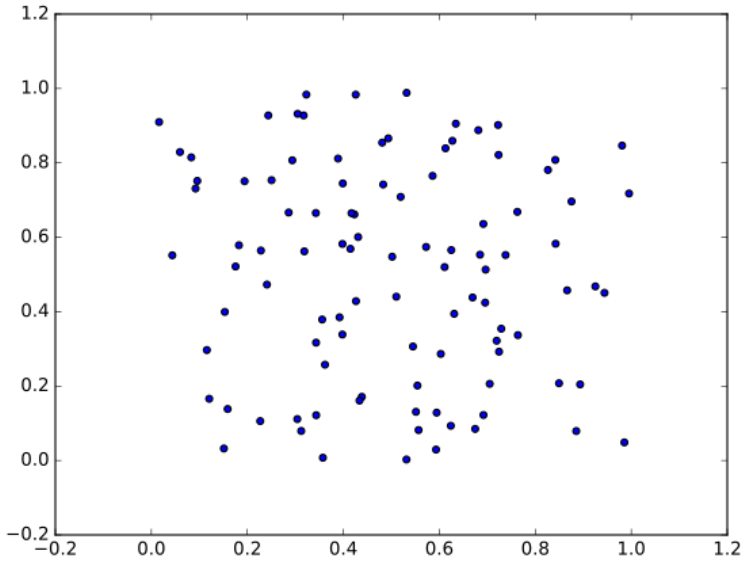
- engl. *bar chart*
- Varijanta stupčastog dijagrama – isto kao stupčasti dijagram, samo položeno
- Izbor između stupčastog i horizontalnog stupčastog dijagrama uglavnom je stvar ukusa

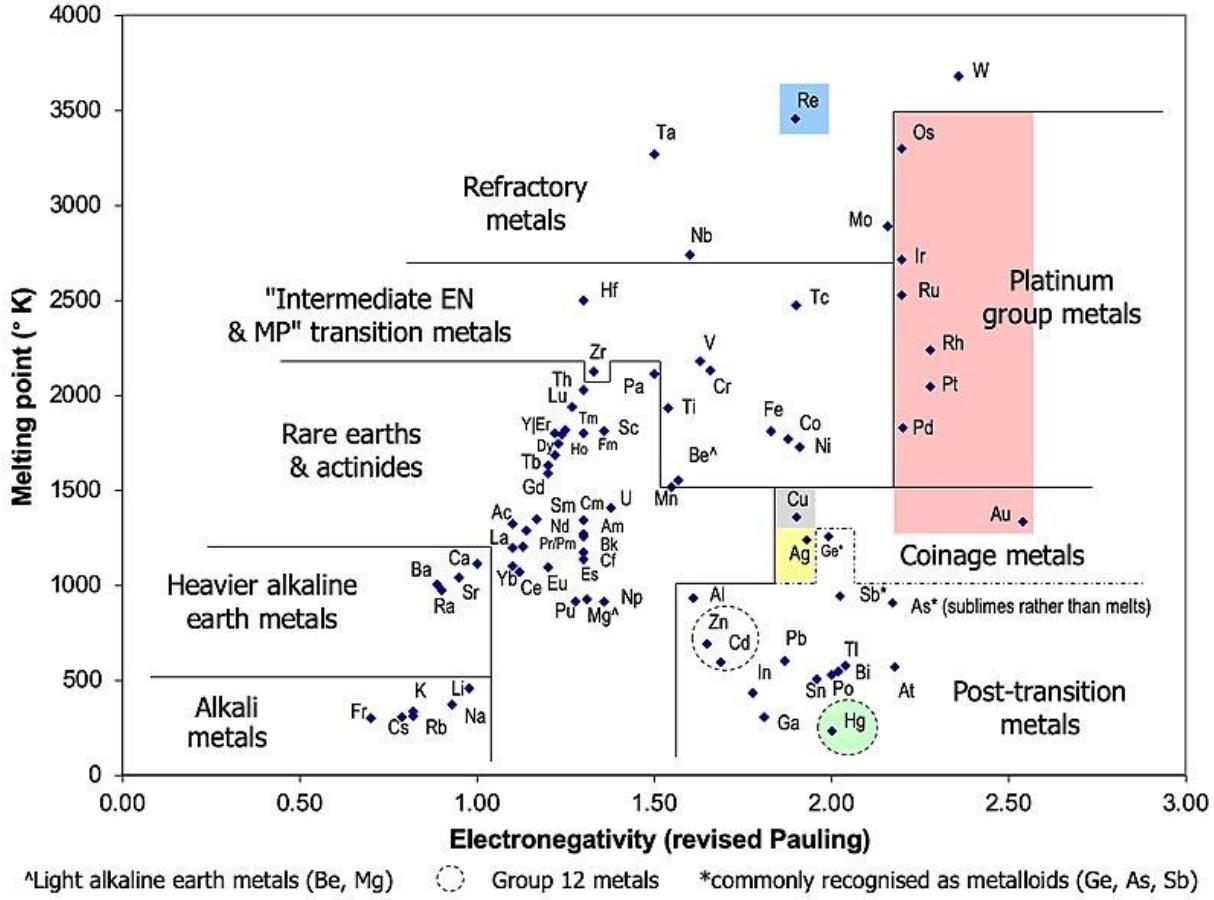


Source: https://commons.wikimedia.org/wiki/Category:Demographic_bar_charts#/media/File:Boyton_occupation_chart_1881.jpg

Dijagram raspršenja

- Prikaz odnosa **dviju numeričkih** varijabli
- Mogu biti jednostavniji i složeniji prikazi
- Pogodno za uočavanje korelacije

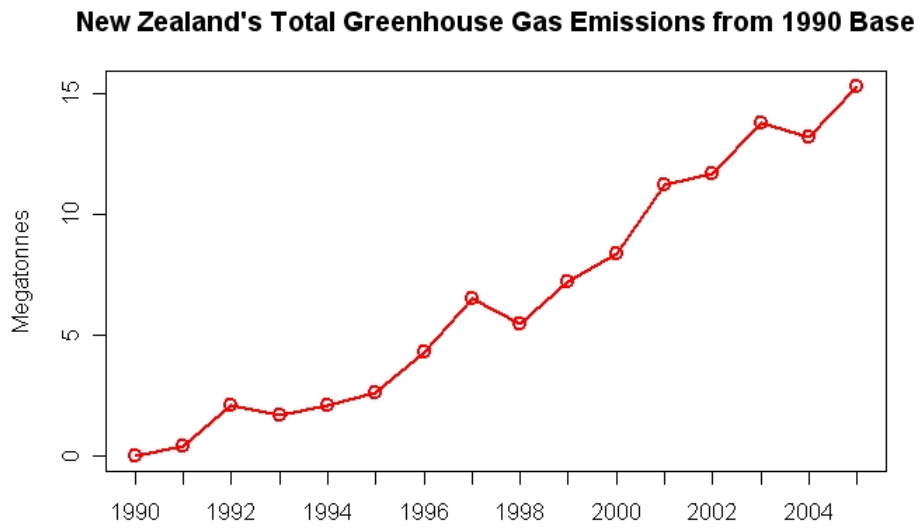


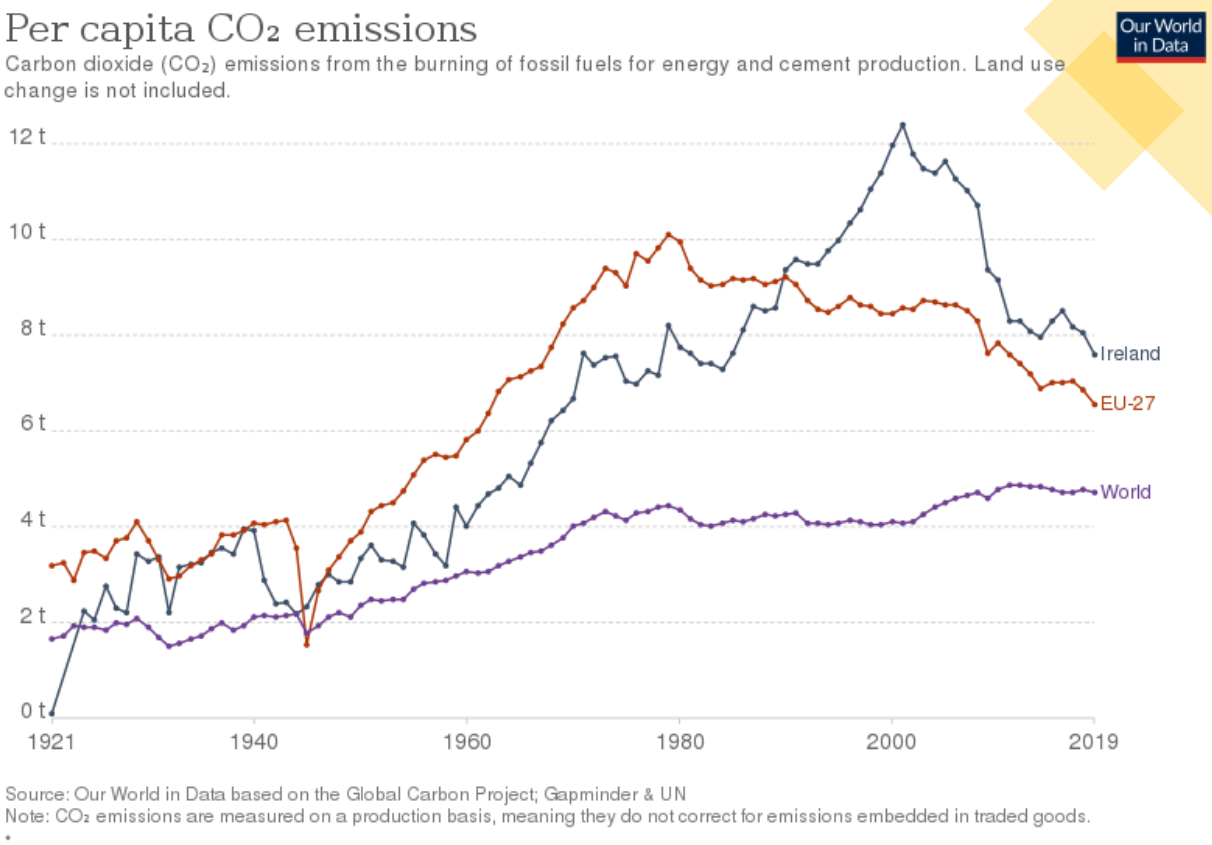


Sources: https://commons.wikimedia.org/wiki/File:Mpl_example_scatter_plot.svg
https://commons.wikimedia.org/wiki/File:Scatter_plot_of_EN_%26_MP_with_NM_shaded.jpg

Linijski dijagram

- Prikaz **jedne ili više numeričkih** varijabli, najčešće u odnosu na varijablu vremena
- Pogodno za razmatranje i analizu vremenskih nizova podataka

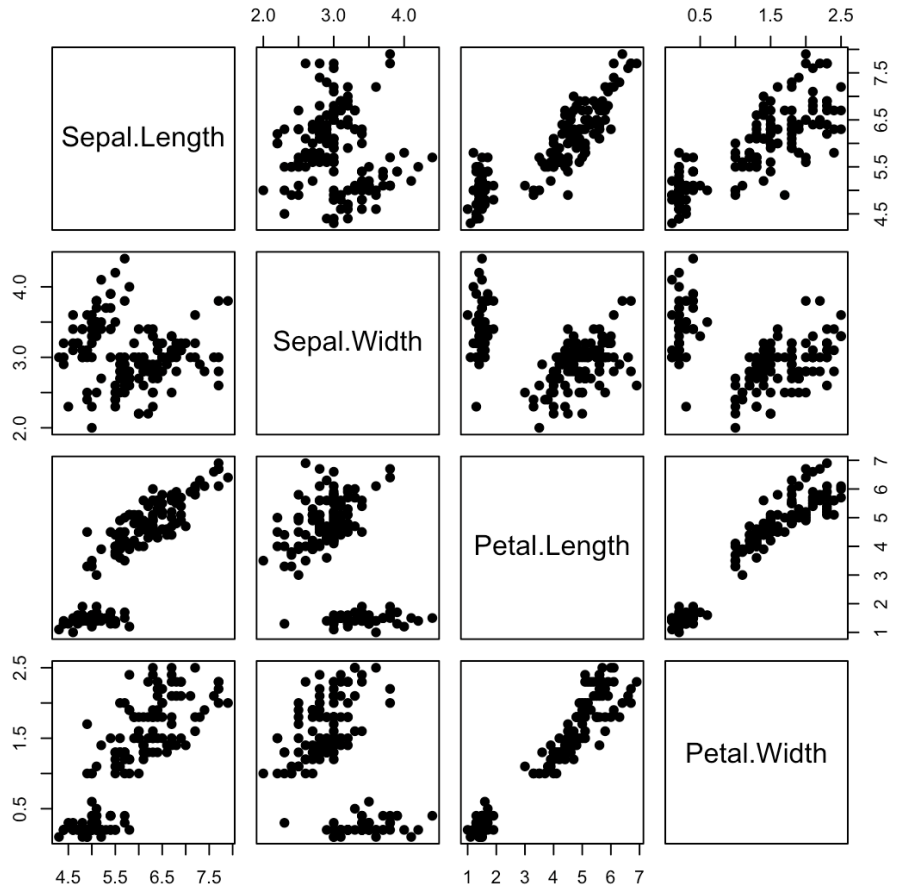




Sources: https://commons.wikimedia.org/wiki/File:New-Zealand-greenhousegases-1990-2005-line-chart.jpeg
https://commons.wikimedia.org/wiki/File:Ireland_v_EU-27_v_World_per_capita_CO2_emissions.svg

Matrica dijagrama raspršenja

- Prikaz odnosa između **više uparenih numeričkih varijabli**
- Korisno za brz pregled mogućih korelacija među varijablama
- Moguće su i varijante matrice dijagrama raspršenja u kojoj se prikazuju i histogrami varijabli te iznosi koeficijenata korelacije na dijagonali i jednoj strani matrice

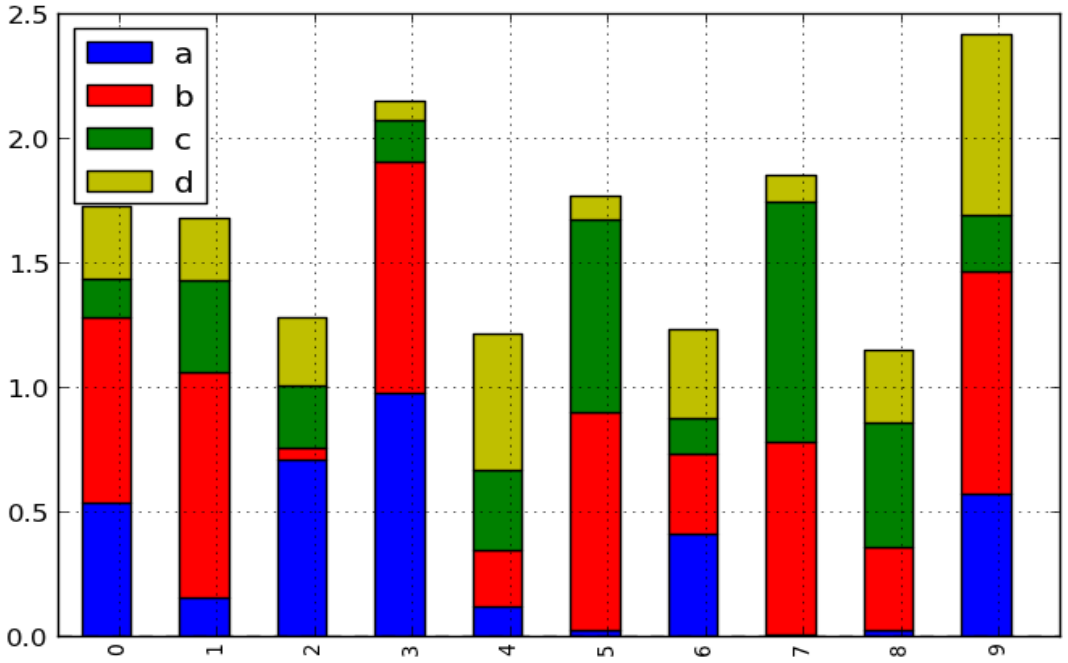


Sources: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

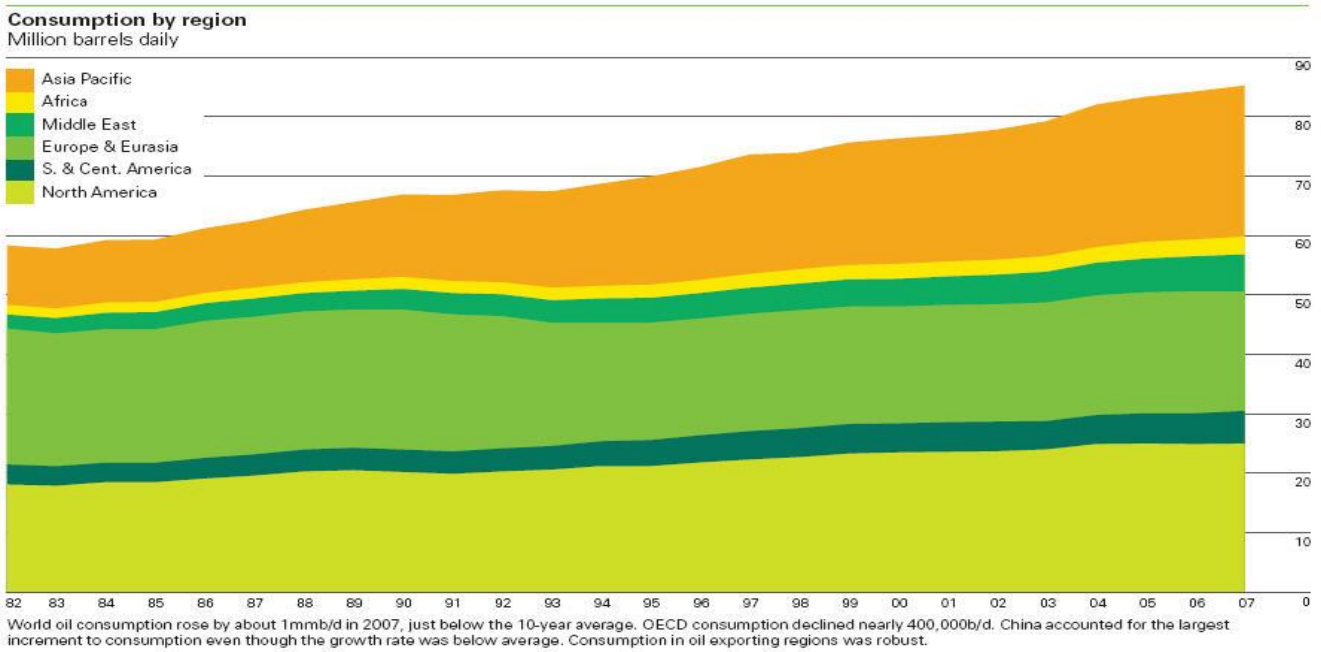
Posloženi dijagram

- Prikazuje međuodnos **tri ili više** varijabli, dolazi u nekoliko varijanti

Dvije kategoričke (0-9, a-d), jedna numerička varijabla (visina stupca)



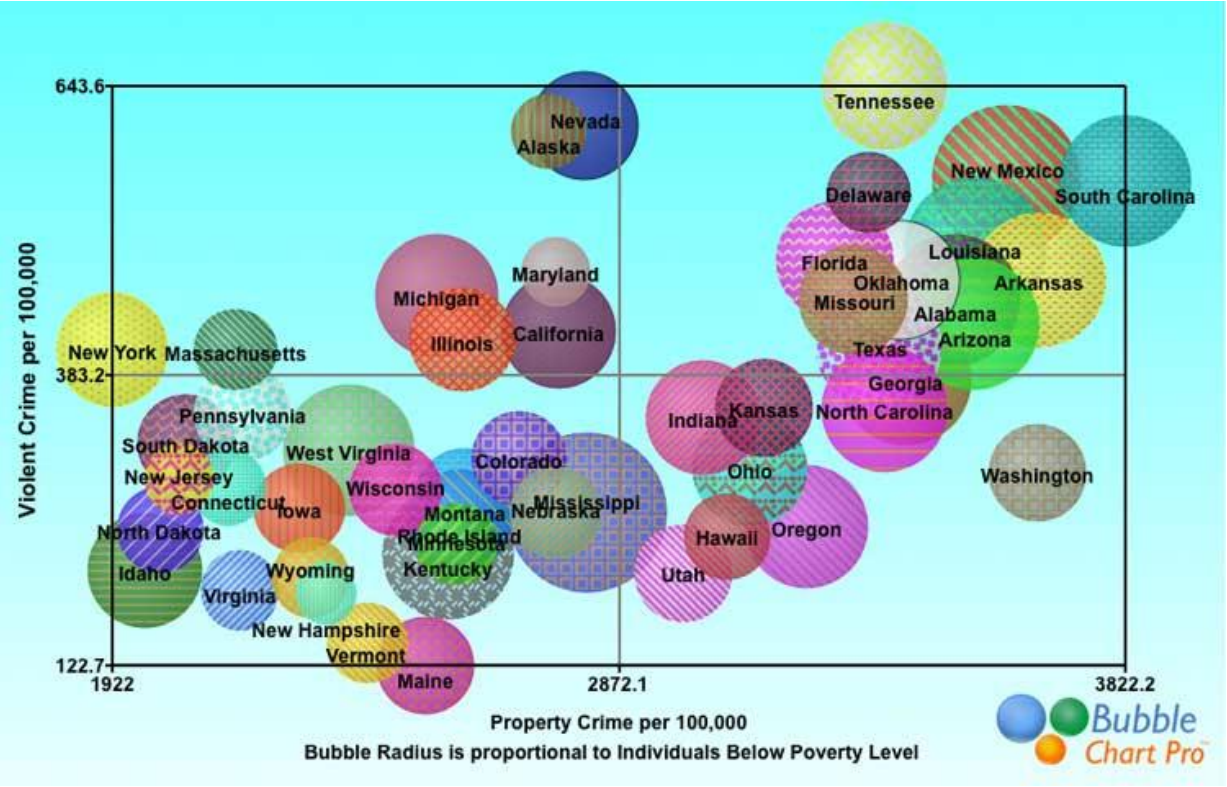
Jedna kategorička (boje kontinenata), dvije numeričke varijable (godine od ‘82. do ’07, iznos potrošnje nafte)



Sources: ADA, 3rd lecture, EPFL, 2020

Mjehuričasti dijagram

- Prikazuje međuodnos **tri** (u 2D) ili **četiri** (u 3D) **numeričkih** varijabli, veličina i/ili boja mjehurića odražava vrijednosti jedne od numeričkih varijabli



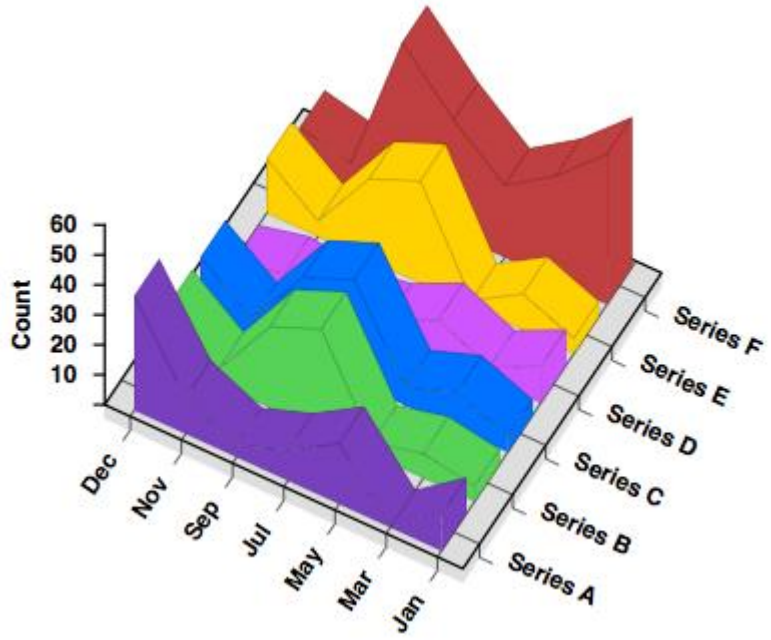
U primjeru:

- Na x-osi: stopa zločina vezanih uz vlasništvo (npr. pljačka kuće)
- Na y-osi: stopa zločina vezanih uz nasilje nad osobom (npr. ranjavanje)
- Mjehurići označavaju postotak građana ispod razine siromaštva, ukratko, veći krugovi – veće siromaštvo
- Jedan primjer = jedna država SAD-a

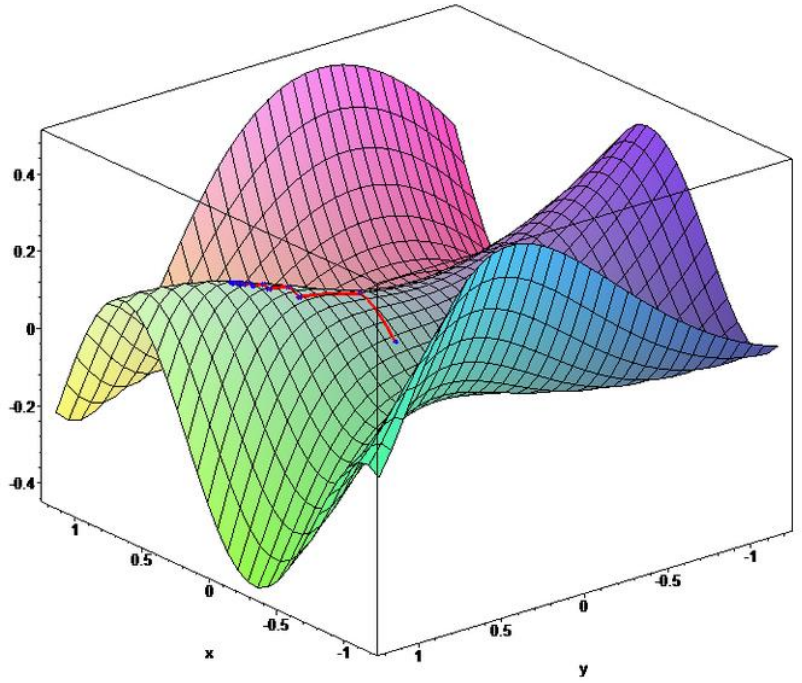
Sources: https://commons.wikimedia.org/wiki/Category:Bubble_charts#/media/File:Bubble_Chart_of_Crime_versus_Poverty_in_50_states.jpg

Površinski 3D dijagram

- Prikazuje međuodnos **tri** ili **četiri numeričkih** varijabli (ako su neke od njih kategoričke, onda se govori o stupčastom 3D dijagramu)



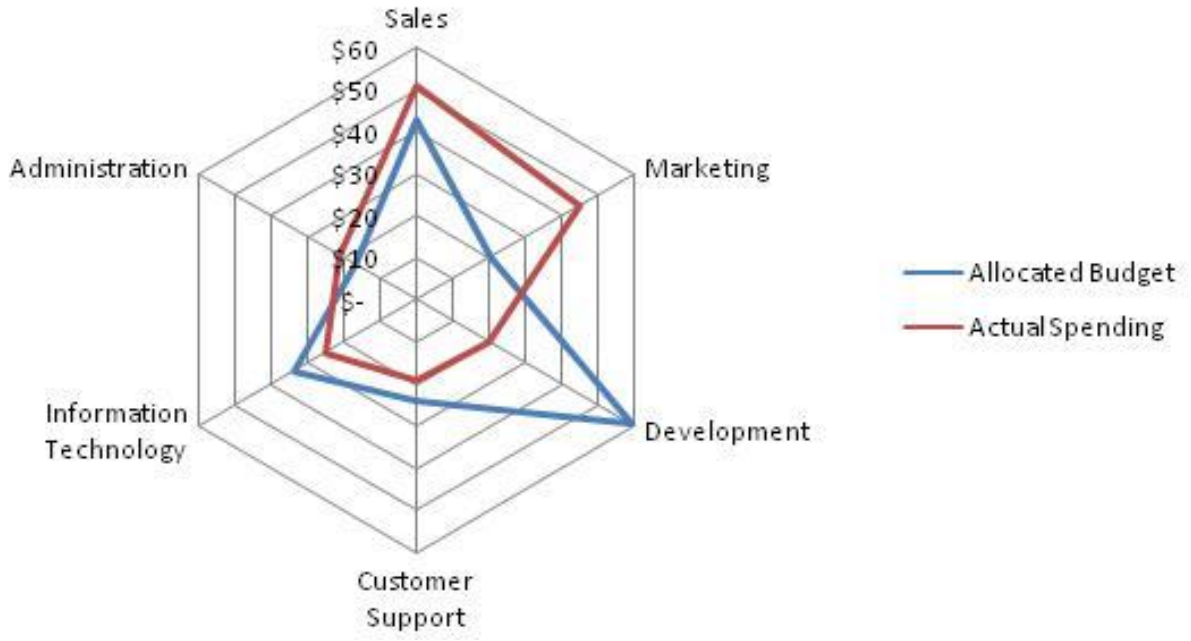
Source: https://www.gigawiz.com/3d-area.html



Source: https://commons.wikimedia.org/wiki/File:Gradient_ascent_(surface).png

Radarski dijagram

- Prikazuje vrijednosti **pet ili više** varijabli određenog primjera (uzorka), više od jednog uzorka prikazuje se radi usporedbe



- U primjeru: dva uzorka – dodijeljeni budžet i stvarna potrošnja, mjerenih prema 6 varijabli

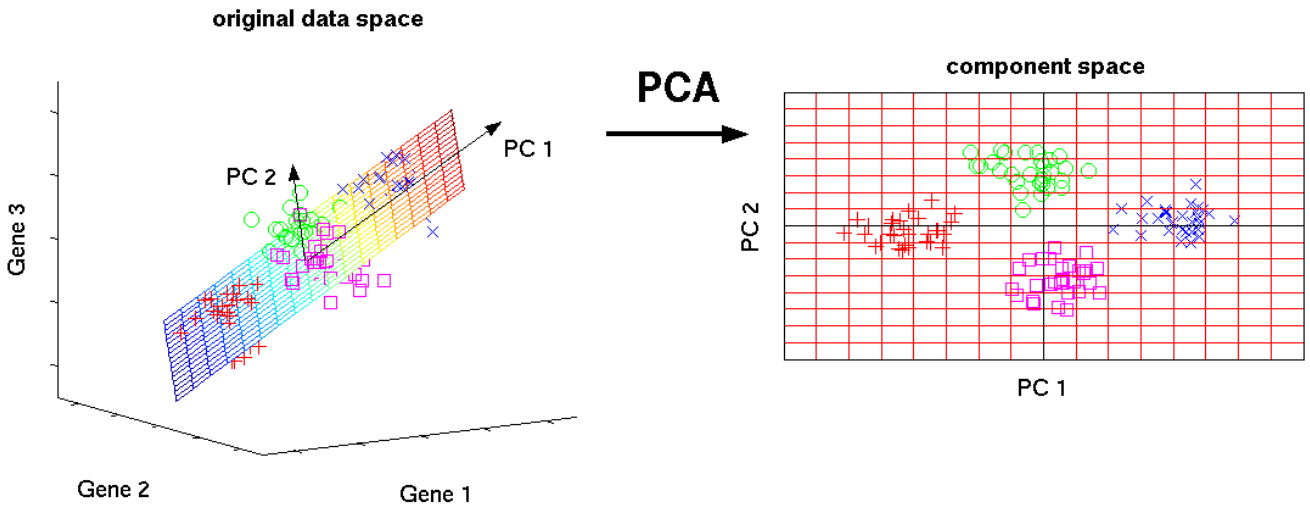
Sources: https://commons.wikimedia.org/wiki/File:Spider_Chart2.jpg

Postupci redukcije dimenzionalnosti s primjenom u vizualizaciji

- Veliki broj postupaka za redukciju dimenzionalnosti, većina primijenjiva za vizualizaciju
- Pretpostavka: redukcijom čuvamo koliko je moguće inicijalnu informaciju, incijalne značajke su **transformirane** (linearno ili nelinearno)
- Ovdje razmatramo samo najčešće korištene postupke za vizualizaciju:
  - **Analiza glavnih komponenti** (engl. *Principal Component Analysis*, PCA)
  - **Samoorganizirajuća mapa** (engl. *Self-Organizing Map*, SOM)
  - **t-SNE** (engl. *t-distributed Stochastic Neighbor Embedding*)

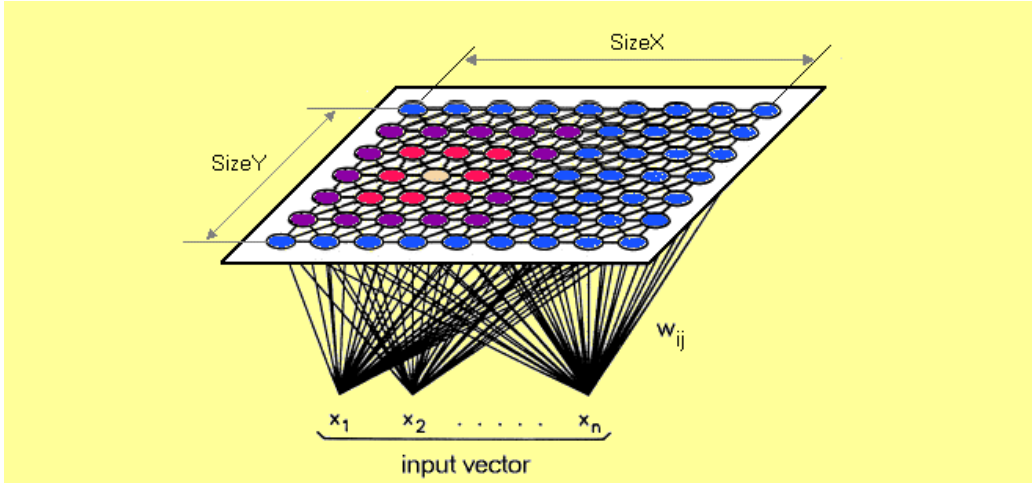
Analiza glavnih komponenti

- 1901., Karl Pearson
- Omogućuje se vizualizacija visokodimenzionalnih numeričkih podataka u niskodimenzionalnom (2D ili 3D) prostoru
- **Glavne komponente**
  - Analitički su izražene kao linearna kombinacija izvornih značajki i to takva da redom pokrivaju **najveću varijabilnost** u podacima
  - Međusobno su ortogonalne
- Obično se prve dvije ili tri glavne komponente zadrže za vizualizaciju
- Korisno za otkrivanje **grupa** podataka

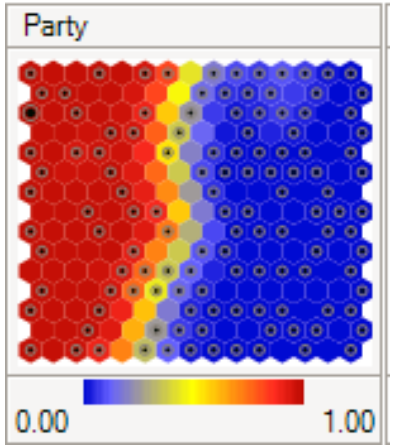


Samoorganizirajuća mapa

- 1982., Teuvo Kohonen
- Umjetna neuronska mreža učena **nenadzirano**, preslikava (**map**ira) ulazne podatke u niskodimenzionalnu (tipično 2D) reprezentaciju
- Težine se uče **kompetitivnim učenjem**, čvorovi se bore za pravo da odgovore na podskup ulaznih podataka, što se dogodi ako su njihove težine najbliže ulaznom vektoru
- U fazi testiranja može odrediti kojoj lokaciji u mapi je novi ulazni podatak najbliži



Source: https://medium.com/@abhinavr8/self-organizing-maps-ff5853a118d4

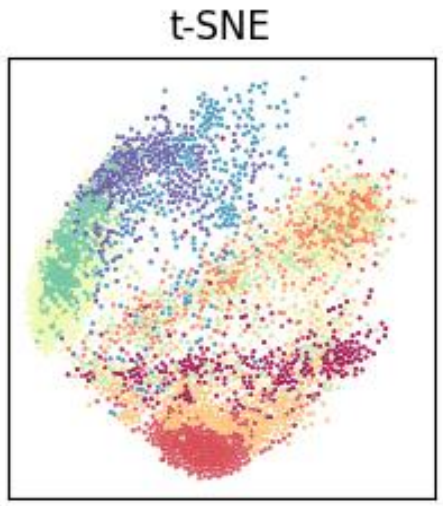


Source: https://en.wikipedia.org/wiki/Self-organizing_map#/media/File:Synapse_Self-Organizing_Map.png

t-SNE

- Tehnika **učenja višestrukosti** (engl. *manifold learning*), Maaten i Hinton 2008.
- Traži niskodimenzionalnu strukturu takvu da svojstva grupiranja u višoj dimenziji ostanu sačuvana
- Srodstvo točaka u visokodimenzionalnom prostoru predstavljeno je Gaussovim zajedničkim vjerojatnostima, a u ugrađenom prostoru Studentovim t-razdiobama (plosnatija je)
- Kullback-Leiblerova divergencija između zajedničkih vjerojatnosti u izvornom prostoru i ugrađenom prostoru se minimizira gradijentnim spustom
- Metoda je računski vrlo zahtjevna ali često daje izvrsno rezultate

Fashion-MNIST: 28x28, 60000 primjera, 10 klasa



Source: Eugen Vušak, „Tehnike učenja višestrukosti za povećanje učinkovitosti analize koja koristi sporedna svojstva kriptografskih uređaja" diplomski rad, FER, 2020.

Ostali grafovi i vizualizacije

- Toplinske mape (engl. *heat map, heatmap*)
  - Odnos 3 ili 4 varijable, razne primjene
  - https://www.optimizely.com/optimization-glossary/heatmap/
- Oblak riječi (engl. *word cloud*)
  - Primjena u vizualizaciji teksta (*bag of words*)
  - https://www.wordclouds.com/
- Graf mreže teksta (engl. *text network graph*)
  - Prikazuje povezanost riječi u tekstu (kontekst)
  - https://infranodus.com/
- ...

Principi i najbolje prakse vizualizacije

Izbor boje i tonova

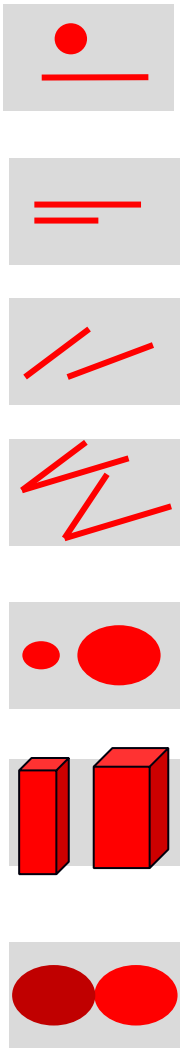
- Kod sivih tonova potrebna je **značajna razlika u intenzitetu** da bi se razlika uočila
  - Izbjegavati „jedva-primjetnu razliku”, vidjeti Weber-Fechnerov zakon: https://en.wikipedia.org/wiki/Weber%E2%80%93Fechner_law
  - Bolje je „igrati na sigurno” i koristiti mali broj jako različitih tonova

Stvarno: kontinuirani (skoro) spektar tonova boje

Ono što ljudsko oko dobro percipira – diskretni (manji) broj različitih tonova

Percepcija magnitude

Najtočnije



Najmanje točno

Pozicija

Duljina

Nagib

Angle

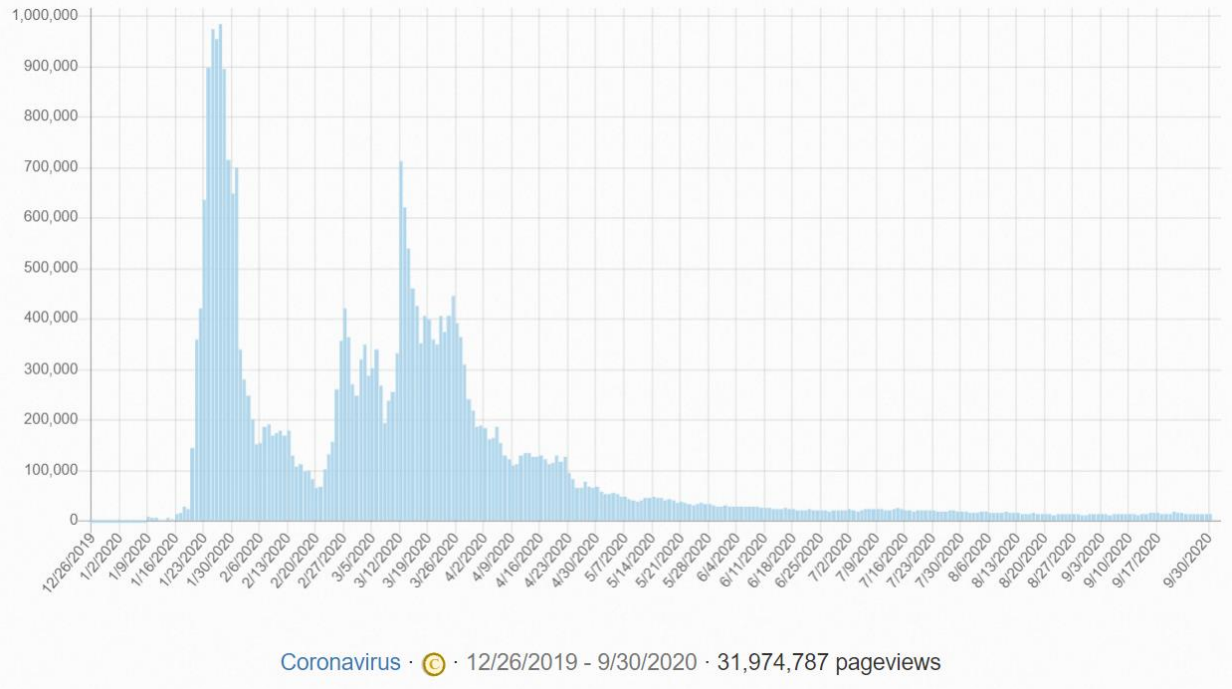
Area

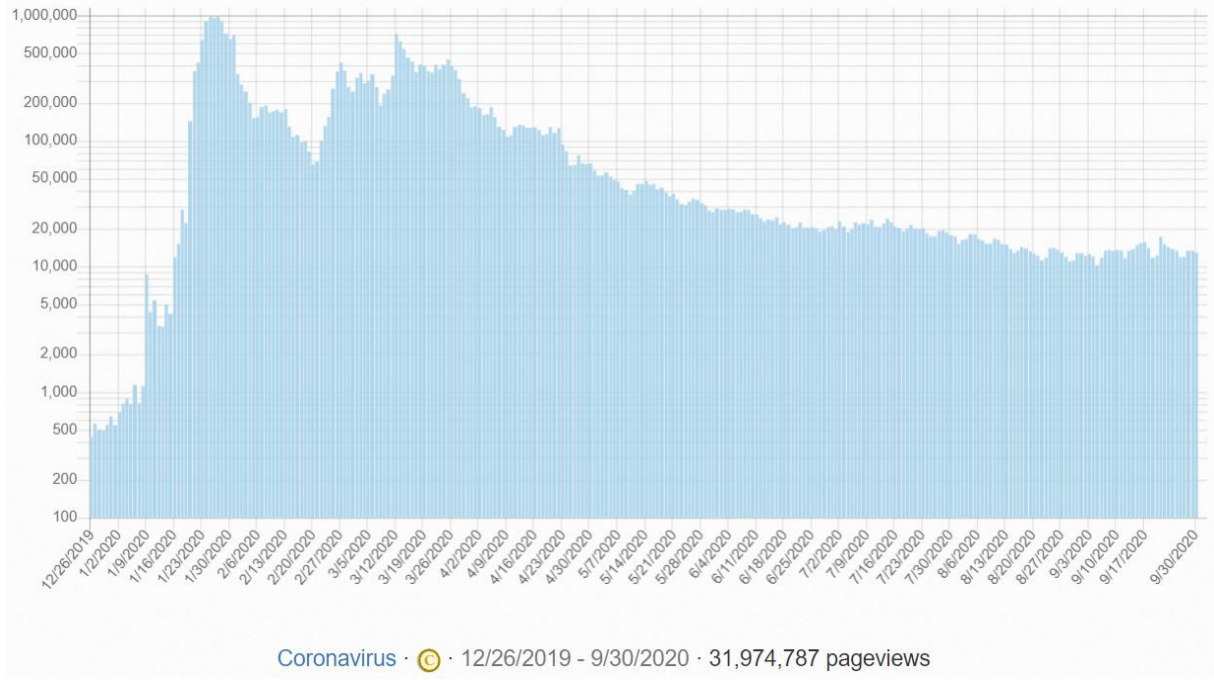
Volume

Ton-zasićenje-gustoća boje

- U prikazu se fokusirati na što češće korištenje elemenata slike s **najtočnijom percepcijom magnitude**
- Izbjegavati male (i nejasne) razlike

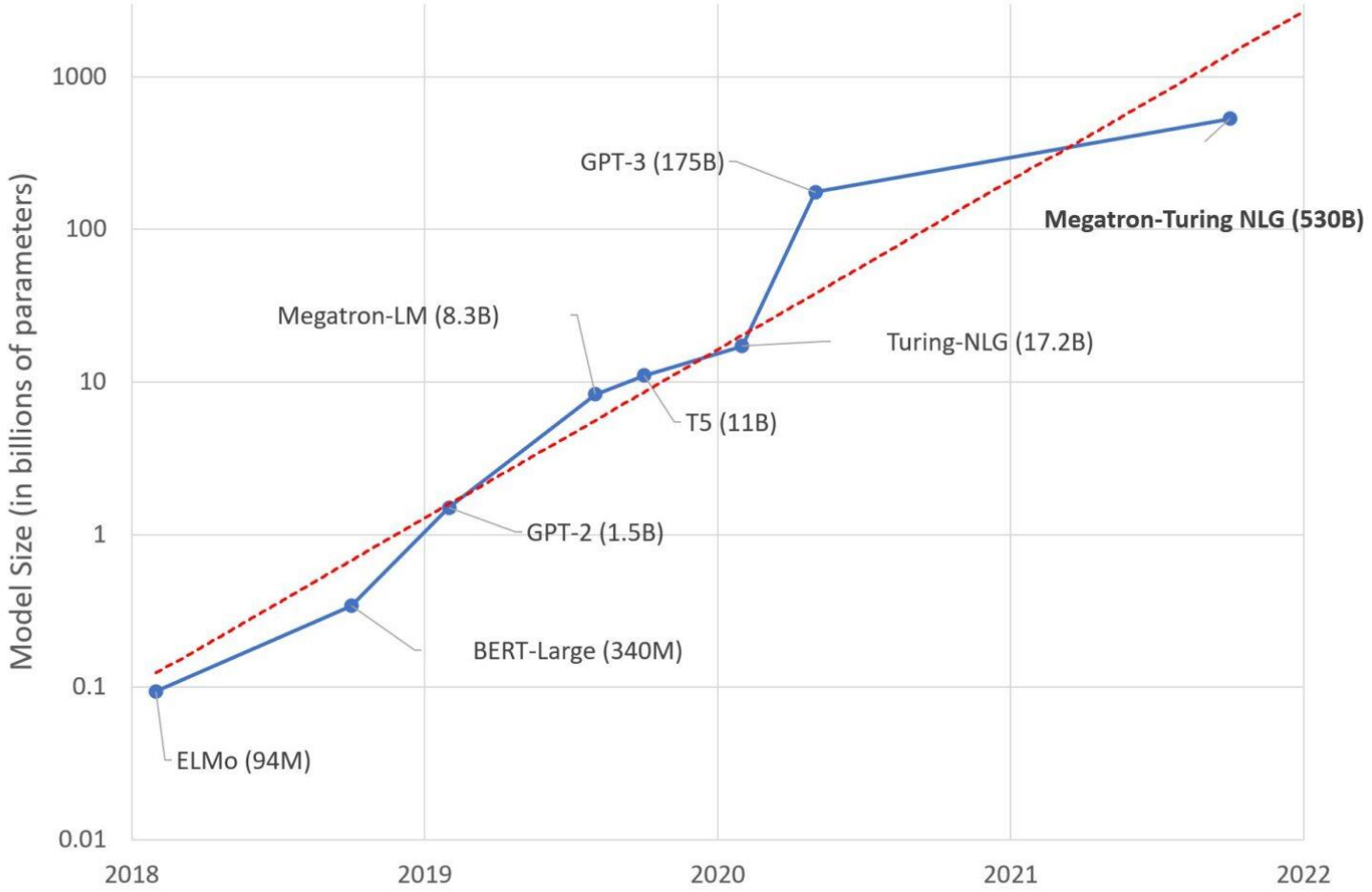
Pripaziti na skale prikaza grafova





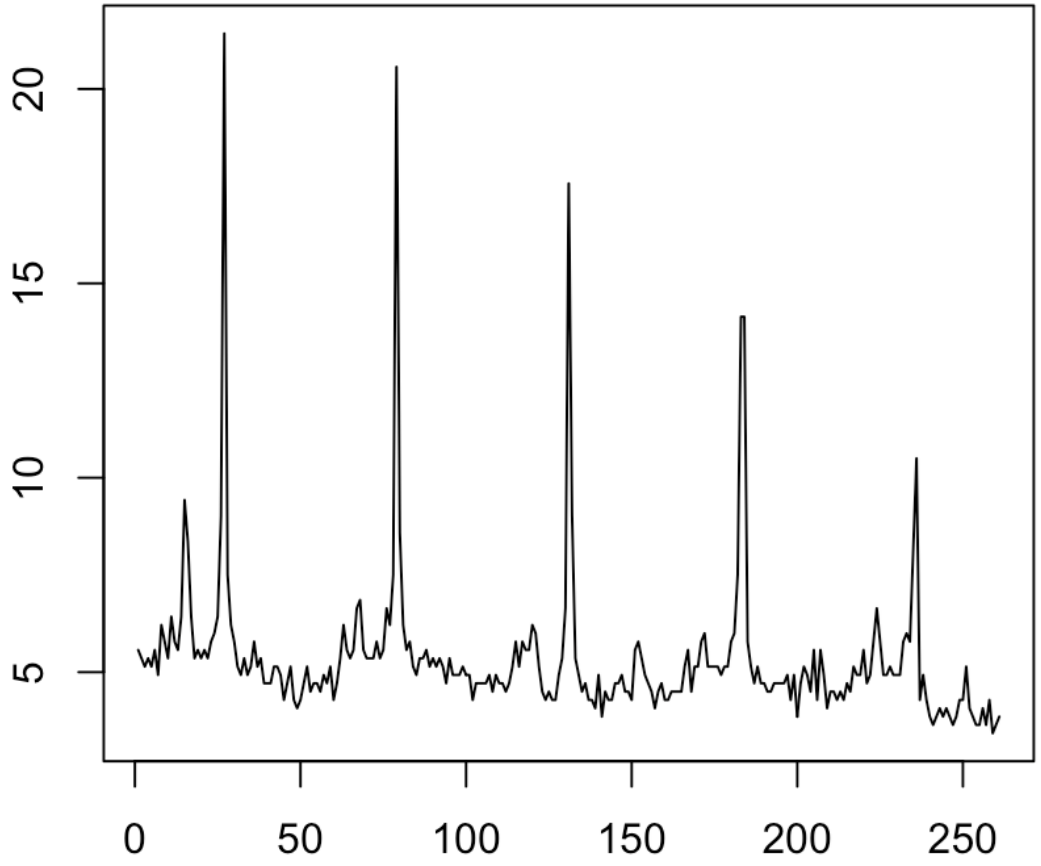
- Neki put je bolja za prijenos informacije uobičajena linearna skala, a neki put logaritmska
- Važno je skalu uočiti na vrijeme!

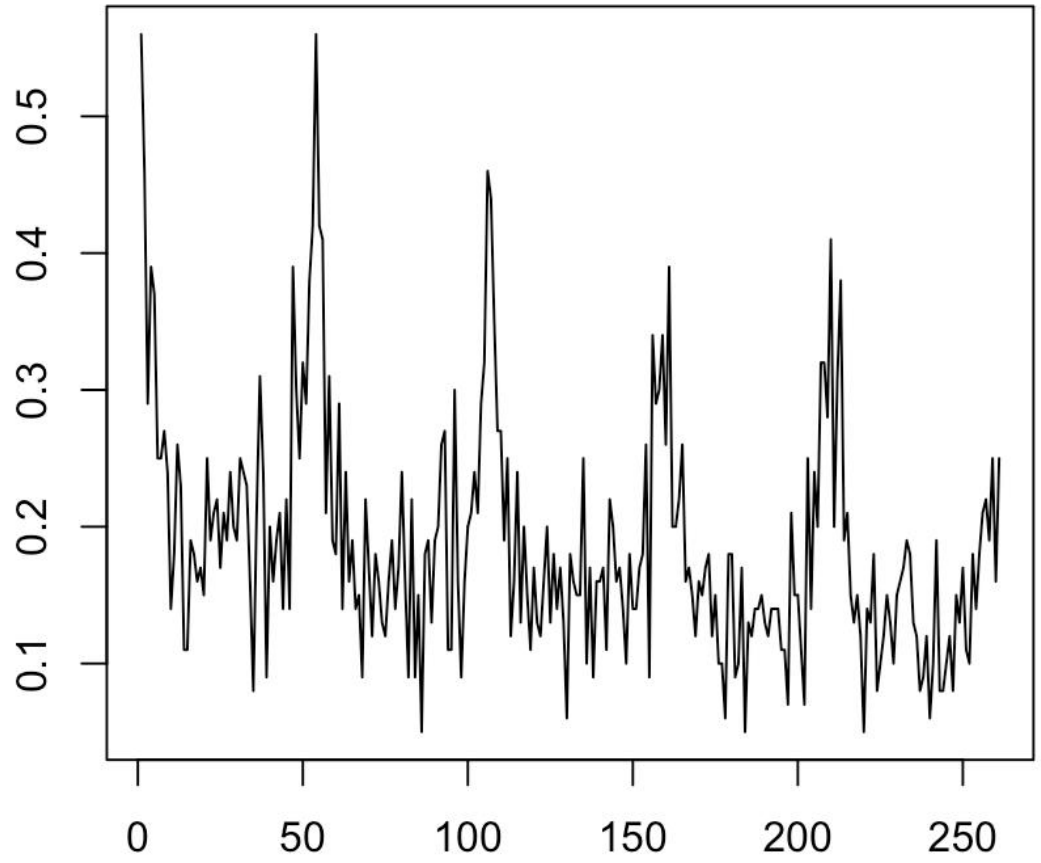
Pripaziti na skale prikaza grafova



Source: https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/
https://towardsdatascience.com/counting-no-of-parameters-in-deep-learning-models-by-hand-8f1716241889

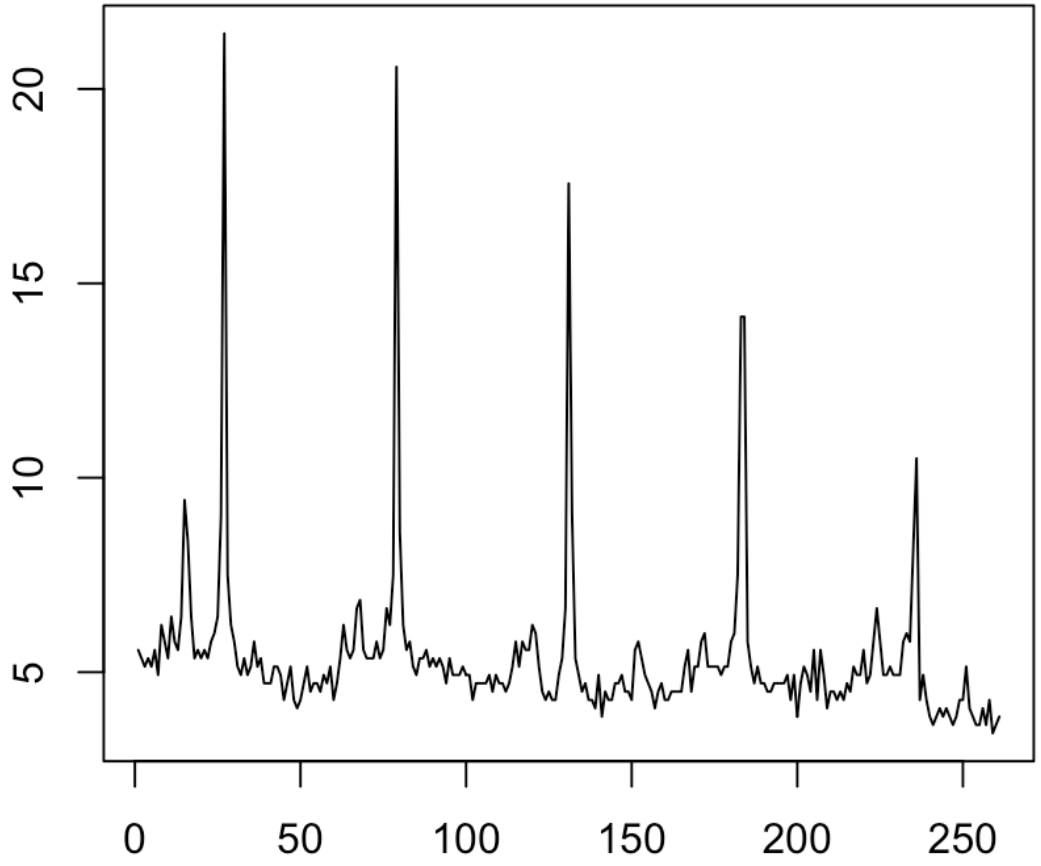
Brzo odgovoriti: koji vremenski niz ima višu srednju vrijednost?

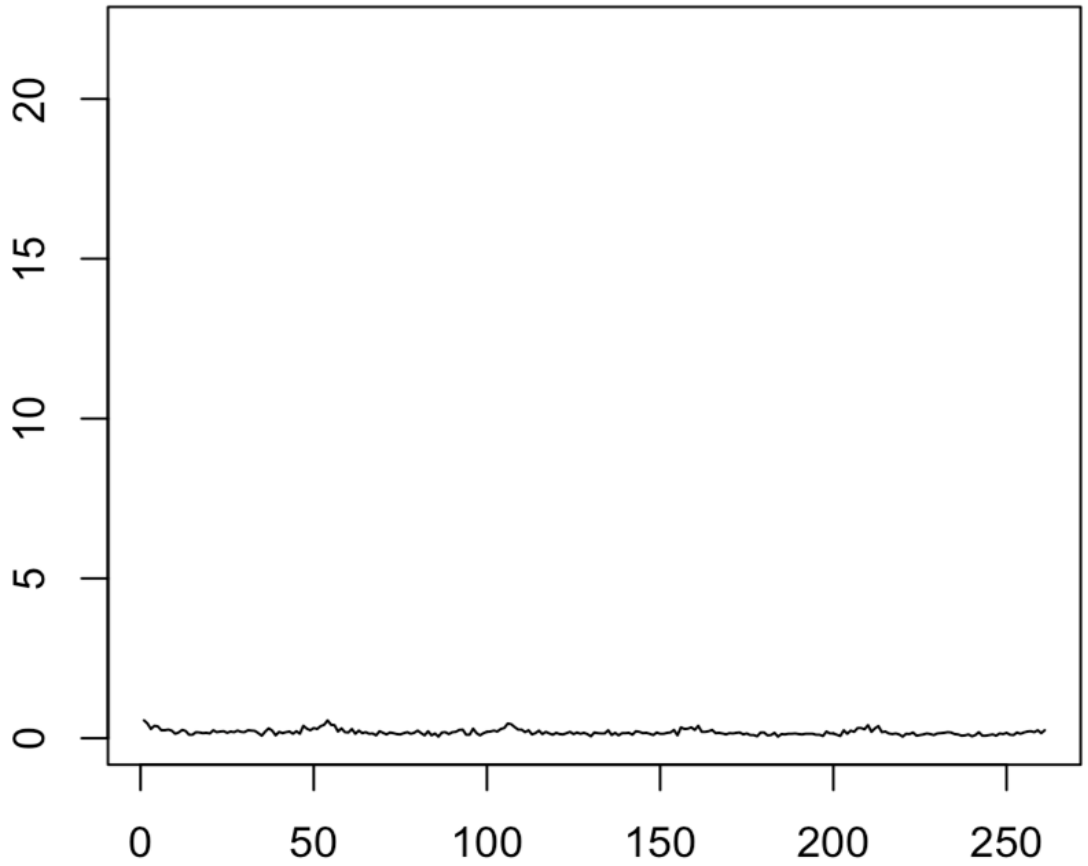




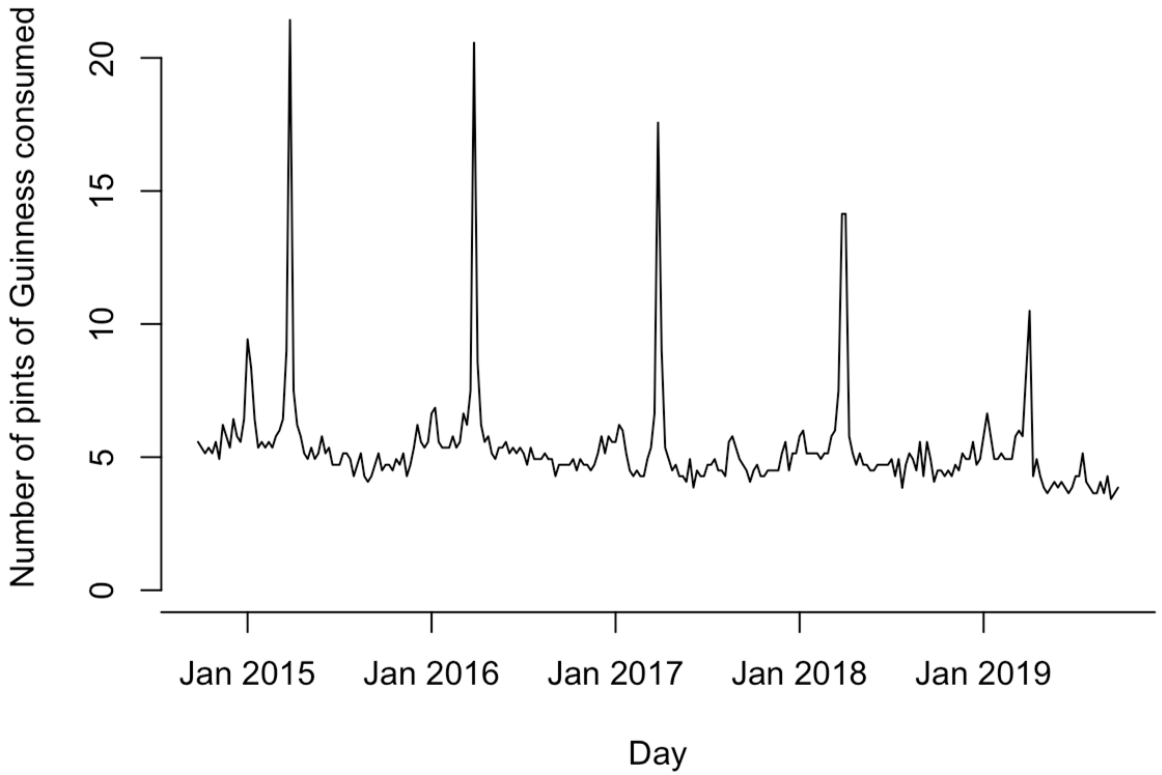
Source: ADA, 3rd lect., EPFL, 2020.

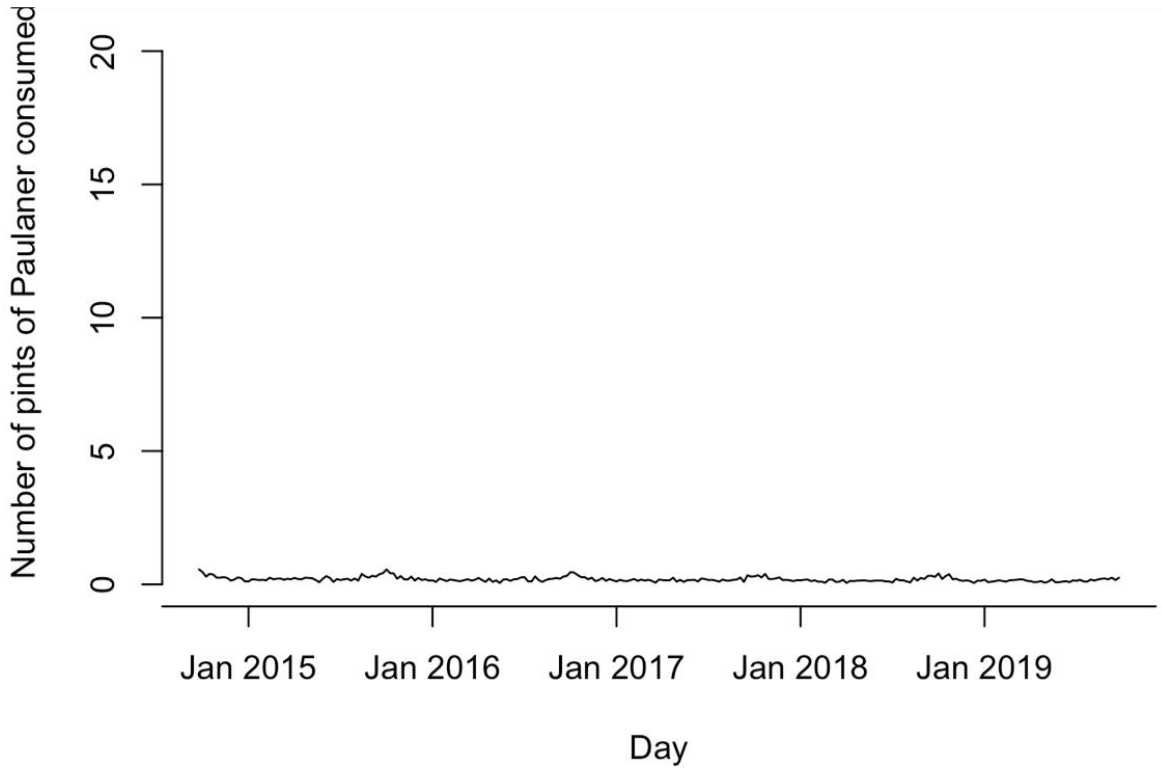
Stvarna situacija



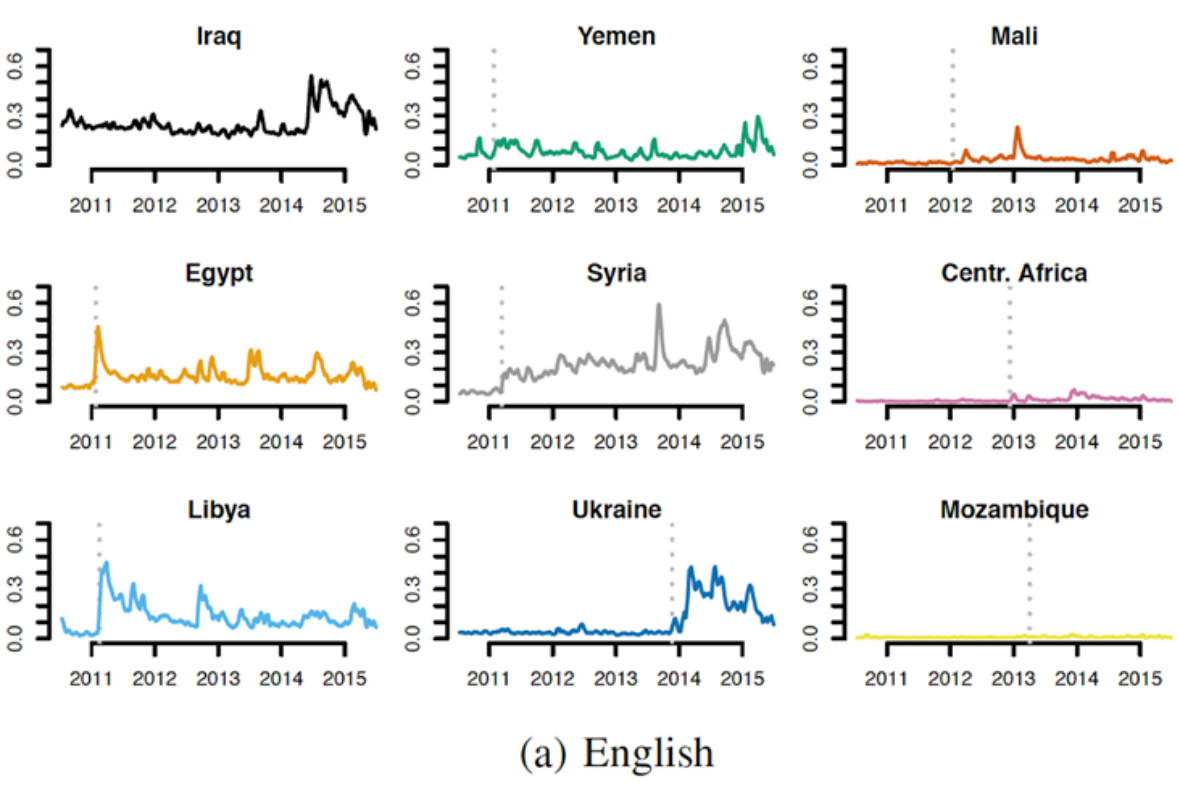


Uvijek označiti osi!

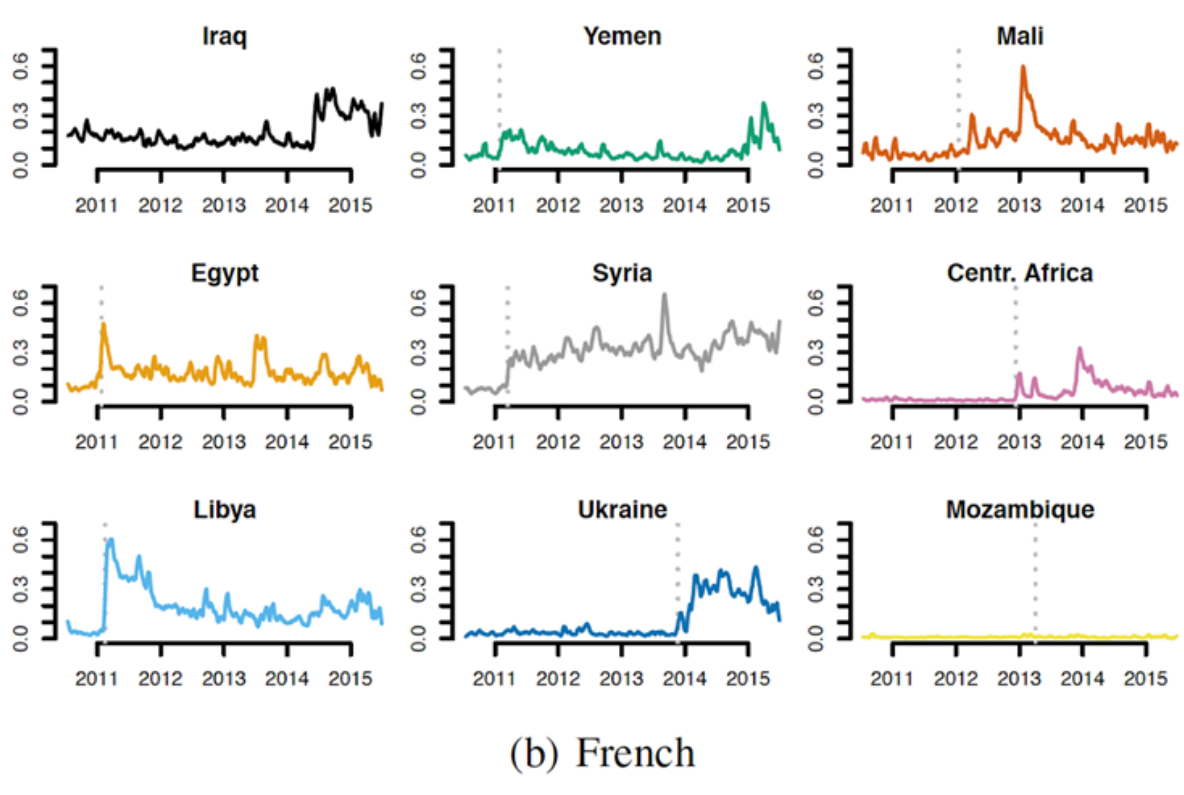




Boje i skale se trebaju koristiti konzistentno između različitih primjera vizualizacije!



(a) English

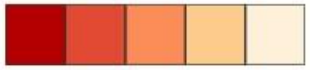


(b) French

Source: ADA, 3rd lect., EPFL, 2020.

Sheme i raspoznavanje boja

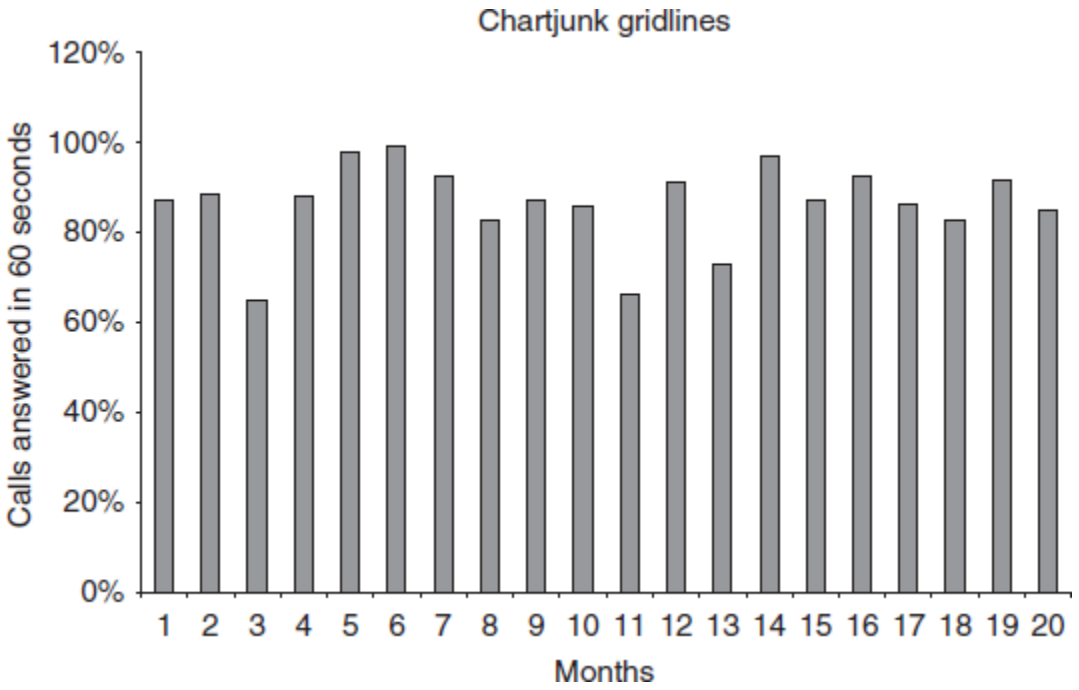
- Različite sheme boja
  - **Sekvencijalna** – za uređene numeričke varijable, niže vrijednosti svijetli tonovi, više vrijednosti tamni tonovi

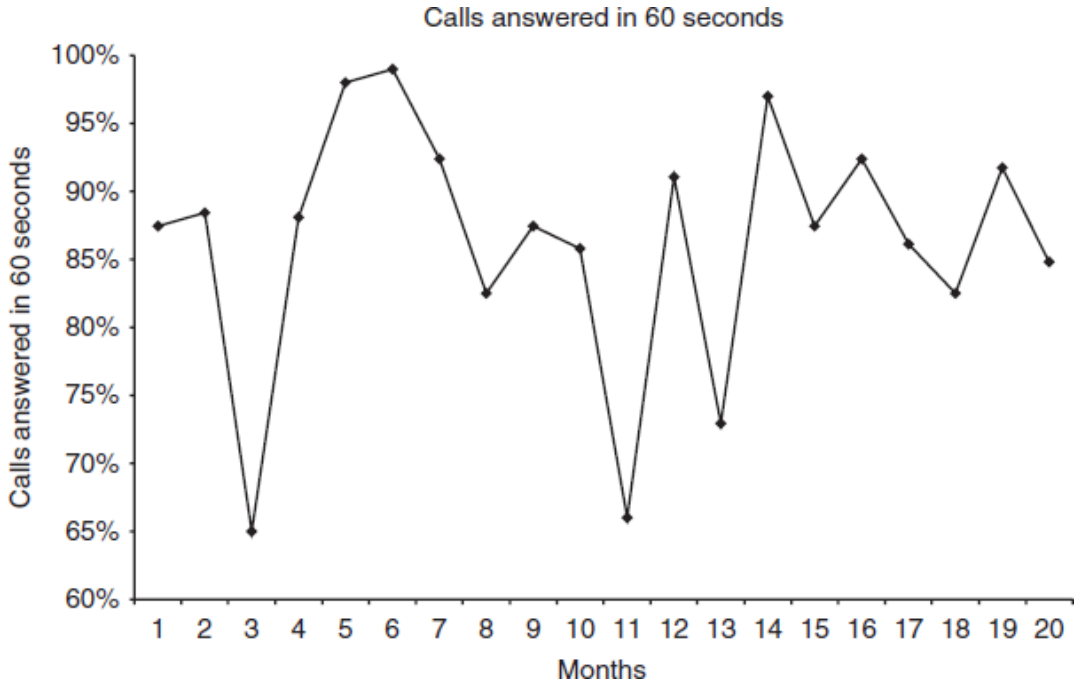
    
  - **Divergentna** – naglasak je na srednjoj vrijednosti s najsvjetlijim tonom, rubne vrijednosti imaju tamne tonove, koristi se za numeričke varijable ako se želi naglasiti srednju vrijednost i rubne vrijednosti

    
  - **Kategorička** (kvalitativna) – koristi se za kategoričke podatke, boje se biraju neovisno o klasi

    
- Koristiti palete prilagođene osobama koje imaju **probleme s raspoznavanjem boja** (engl. *colorblind-safe color palette*)
- Oko 8% muškaraca ima neki oblik problema raspoznavanja boja (samo oko 0.4% žena)
- Pogledati stranicu: https://colorbrewer2.org/

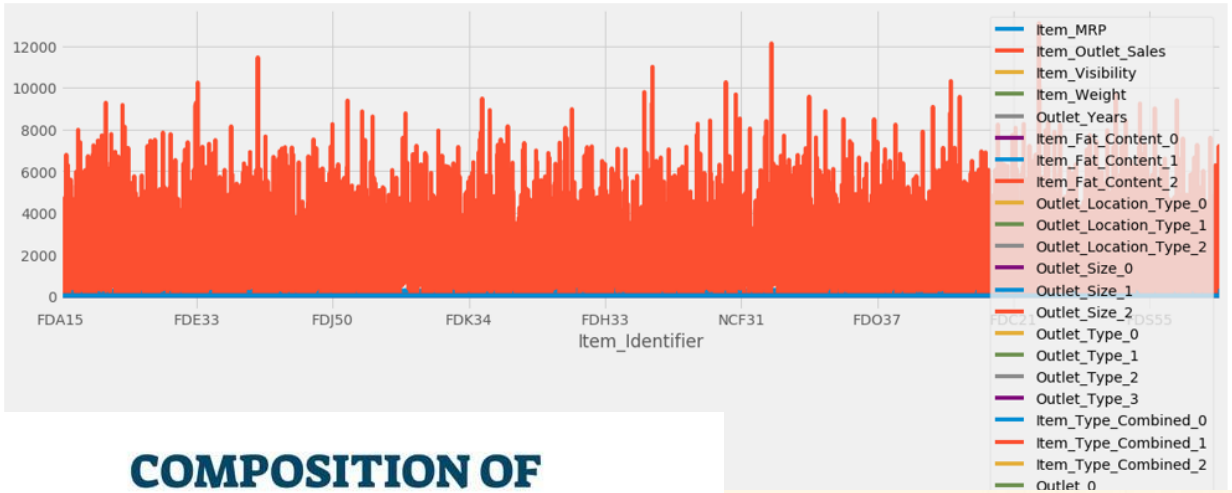
Paziti na izbor grafa, optimirati „potrošnju” boje



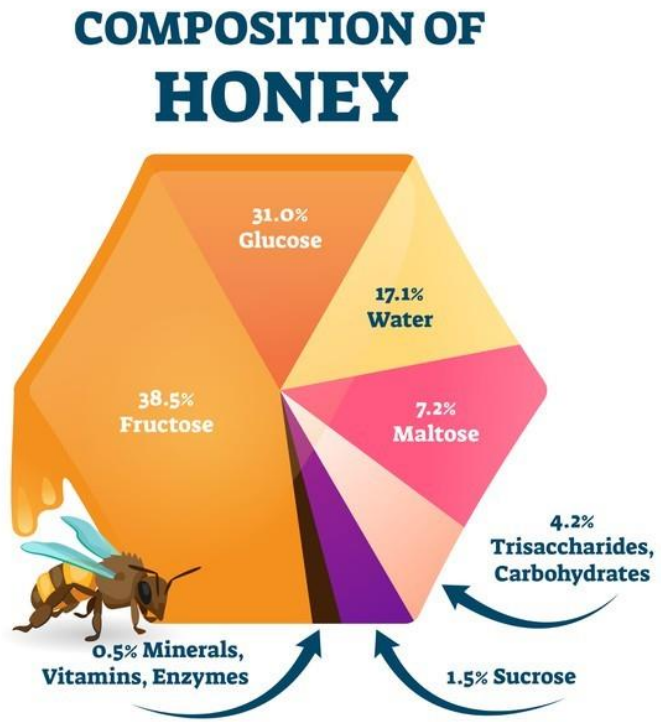


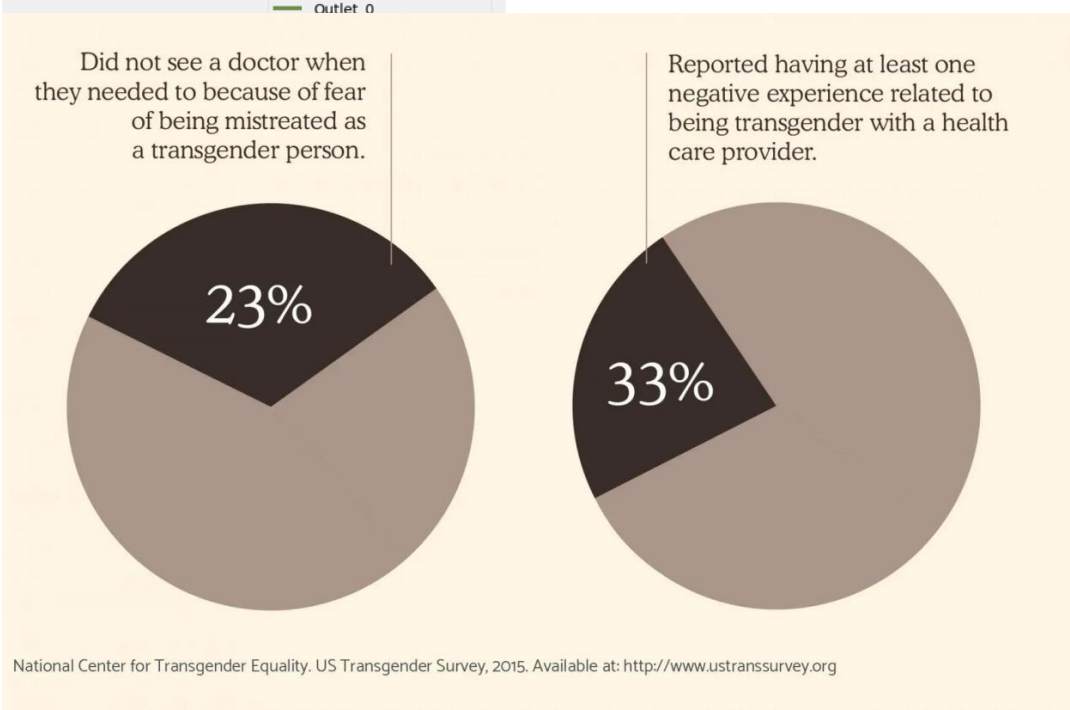
Source: https://www.accessengineeringlibrary.com/content/book/9780071749091/chapter/chapter4

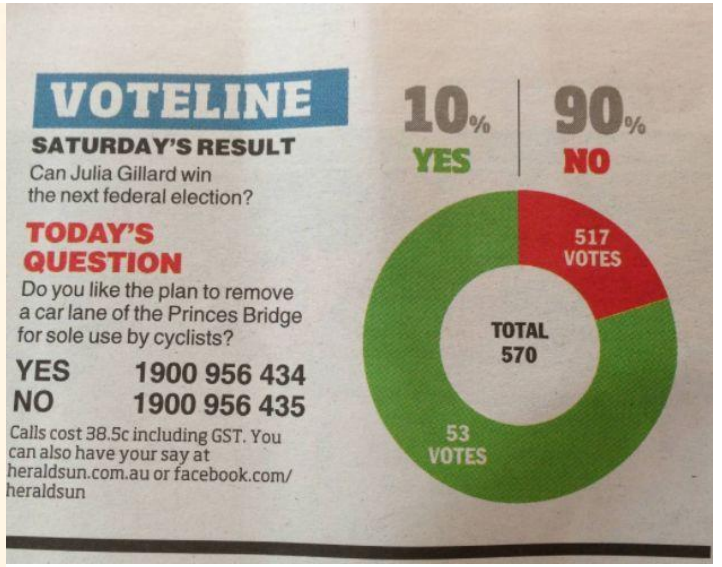
Problematične vizualizacije u praksi: viz.wtf



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Salaried Personel		Jan	Feb	Mar	Apr	may	jun	july	Aug	Sept	Oct	Nov	Dec
2	Utilized 2020	2	24 hrs											
3		1		12hrs										
4		2			36 hrs									
5		4				72 hra								
6		4					76 hrs							
7		3						48hrs						
8		4							72 hrs					
9		2								48 hrs				
10		2									76hrs			
11		4										88 hrs		
12		3											76 hrs	
13		1												48 hrs
14														
15	2021	2	36 hrs											
16		4		84 hrs										



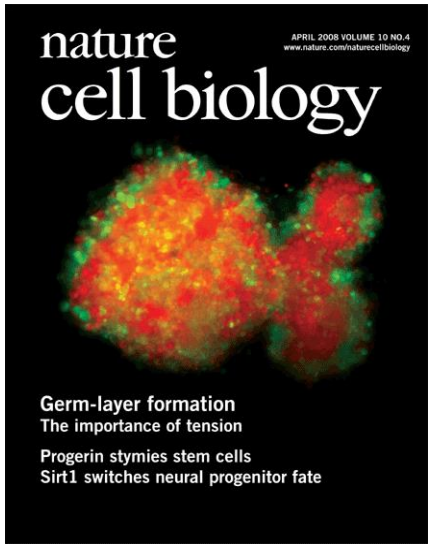


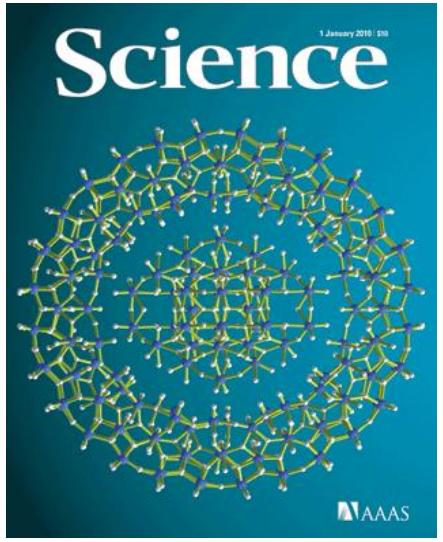


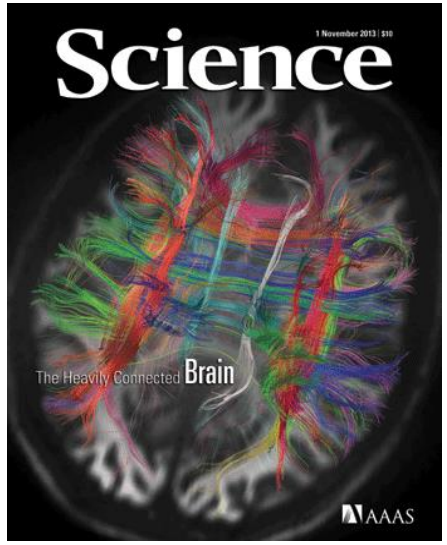
Primjeri korištenja vizualizacije

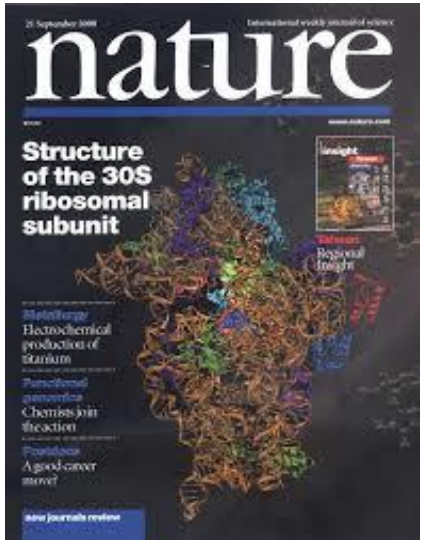
Vizualizacija znanstvenih rezultata

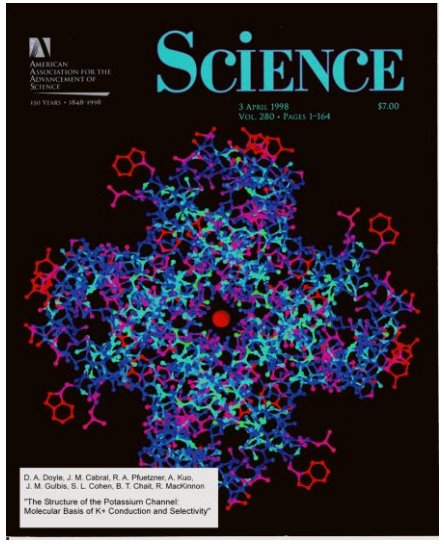
- Često je teško raspoznati što je tu znanost, a što umjetnost ☺
- Primjenjuje sve principe i dobre prakse koje smo diskutirali (i još više)
- Kvalitetne vizualizacije pospješuju čitanost znanstvenih radova



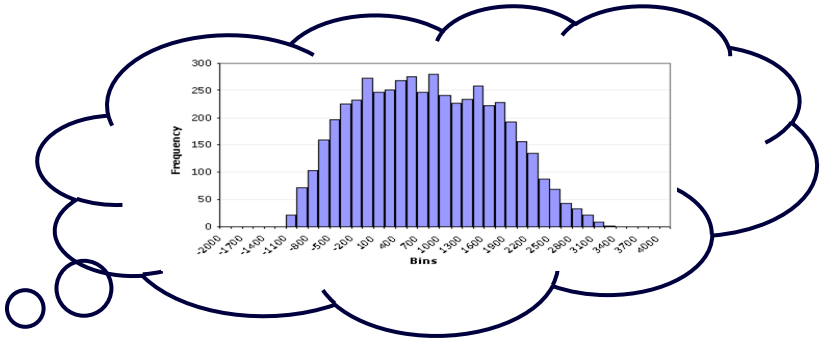




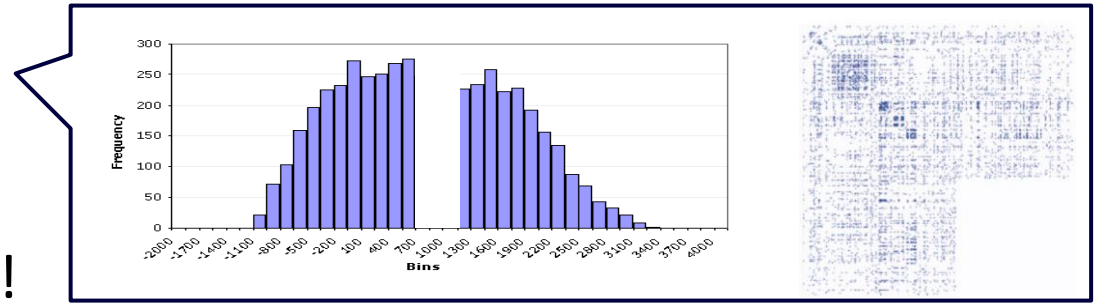




Čudni podaci



- Potrebno je imati teoriju oko toga kako podaci trebaju izgledati
- Neke podatke je vrlo teško objasniti
- Nikada ne zanemariti čudne pojave u podacima, uvijek ih dobro razmotriti!

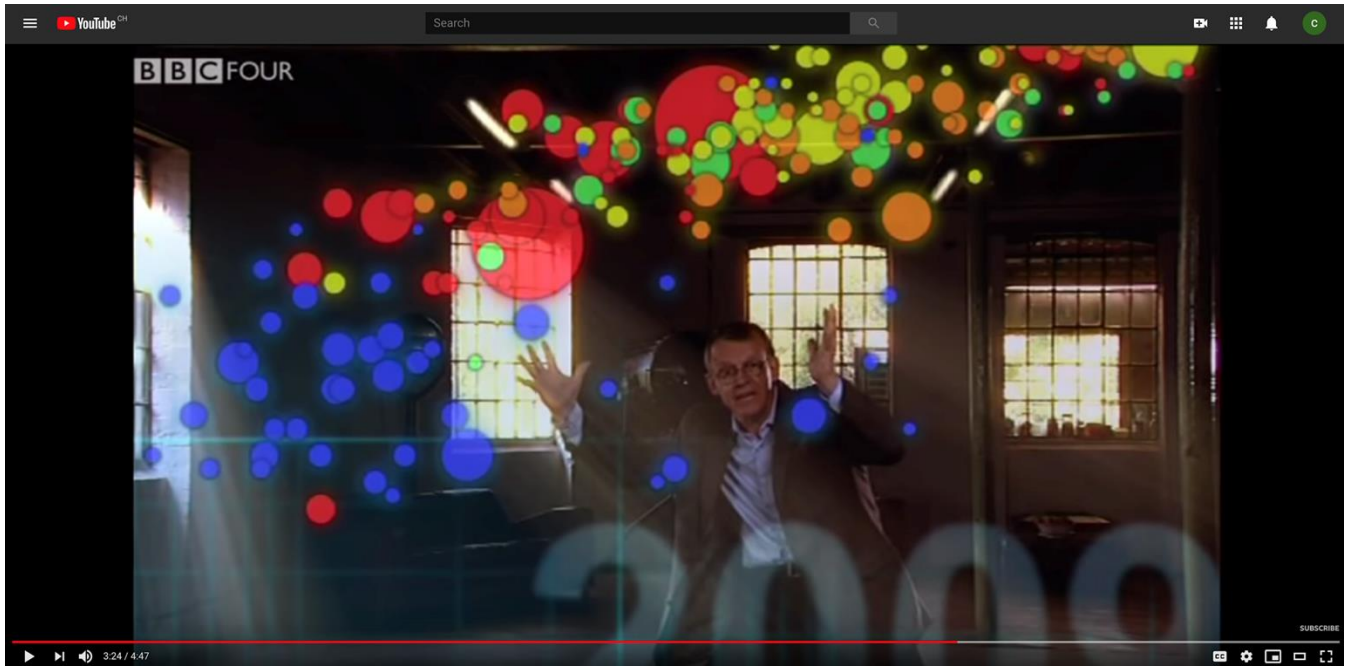


- Ako su vizualizacije dobivene kao rezultat programa, prvo treba pretpostaviti da se radi o programskoj pogrešci, pokušajte ju ispraviti
- Ako nema pogreške, moguće je da se radi o zanimljivom otkriću ☺

Vremenska progresija mjehuričastog grafa

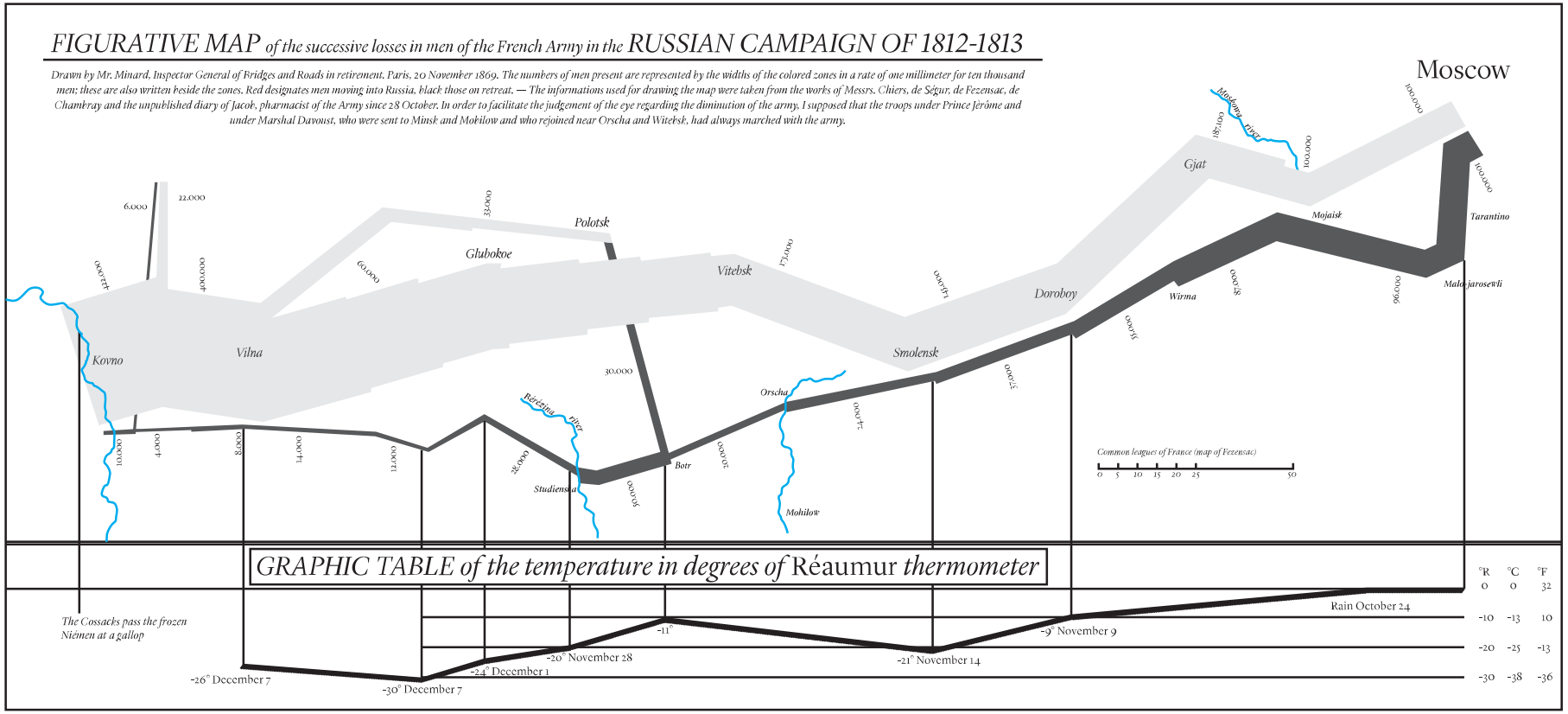
- Hans Rosling: vizualizacija očekivane životne dobi u odnosu na BDP, **200 zemalja svijeta, 200 godina, 4 minute**

- https://www.youtube.com/watch?v=jbkSRLYSojo



Kompliciranе vizualizacije

- Charles Joseph Minard, 1869: Napoleonov pohod



Source: https://en.wikipedia.org/wiki/Charles_Joseph_Minard#/media/File:Redrawing_of_Minard's_Napoleon_map.svg

- Prema Edwardu Tufteu: “Moguće je da se radi o najboljem statističkom prikazu ikad nacrtanom.”
- 5 varijabli: veličina vojske, lokacije, datumi, smjerovi, temperatura zraka pri povlačenju
- https://www.edwardtufte.com/tufte/

Alati za vizualizaciju podataka

D3

- Jedan od najkorištenijih radnih okvira za vizualizaciju

- https://d3js.org/

- Namjerno niskog nivoa, implementiran u JavaScriptu
- Omogućuje najširi raspon opcija crtanja grafova
- Prisutan već dugo, ali i dalje popularan

Vega

- Vega je vizualizacijska gramatika razvijena iznad D3
- Specificira grafiku u JSON formatu
- https://vega.github.io/vega/about/vega-and-d3/
- https://github.com/vega/vega

Vincent

- Vincent je translator Pythona u Vegu (Vincent Vega ☺)
- Značajno olakšava web vizualizacije
- "The data capabilities of Python. The visualization capabilities of JavaScript".
- https://vincent.readthedocs.io/en/latest/

Dash (Plotly)

- Radni okvir niske razine za izradu analitičkih aplikacija u Pythonu
- Izgrađen iznad Plotly.js i React.js, usporediv s D3
- Podržava veliki broj vizualizacijskih mogućnosti kroz sučelja za rad s podacima
- Enterprise Dash je komercijalna verzija koja uključuje razvoj koda, postavljanje u okolinu i integraciju s poslovnom stranom
- https://dash.plotly.com/introduction

Matplotlib

- Opsežna Pythonova biblioteka za izradu statičkih, animiranih i interaktivnih vizualizacija
- Sučelje niskog nivoa
- I dalje jedan od najboljih izbora za vizualizaciju podataka u Pythonu
- https://matplotlib.org/
- https://github.com/matplotlib/cheatsheets

Seaborn

- Pythonova biblioteka za statističku vizualizacija izgrađena iznad Matplotliba
- Sučelje relativno visokog nivoa
- Dosta popularan alat, odličan za većinu jednostavnih vizualizacija
- https://seaborn.pydata.org/

Bokeh

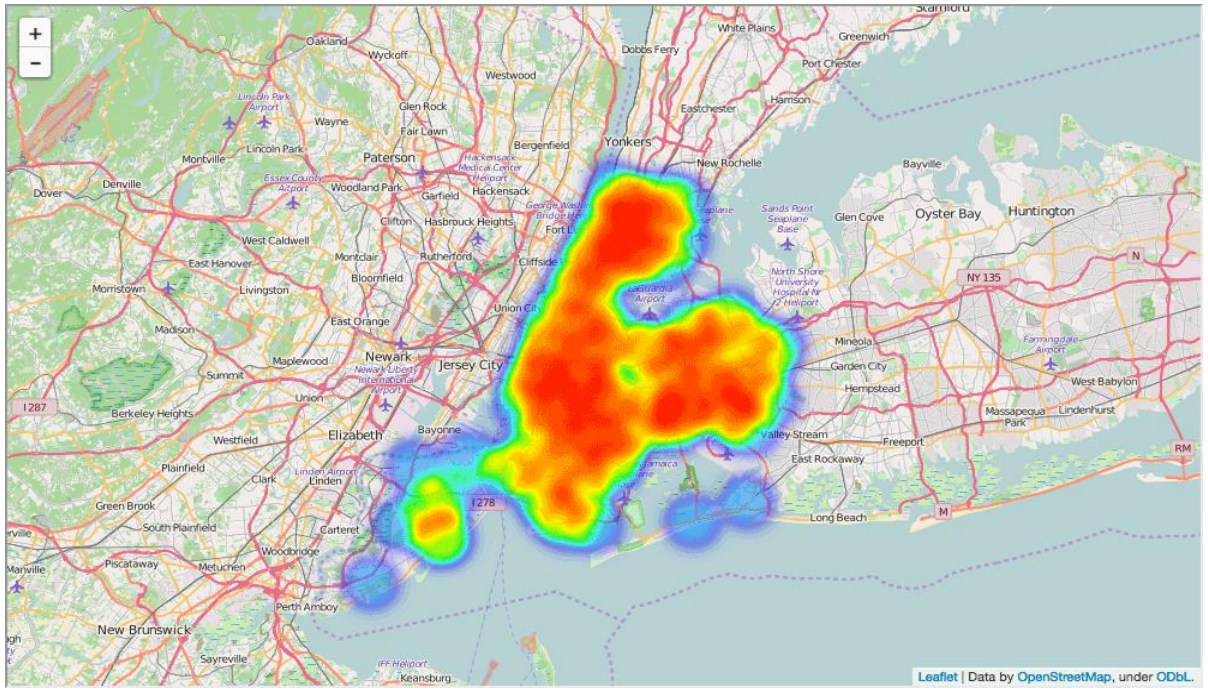
- Neovisna biblioteka za vizualizaciju
- Fokus na vizualizaciji za velike podatke i znanstvene svrhe
- https://bokeh.org/
- https://github.com/bokeh/bokeh

Panel

- Alat u Pythonu za izradu interaktivnih web aplikacija i nadzornih ploča
- https://panel.holoviz.org/index.html

Folium

- Alat za vizualizaciju geopodataka
- https://python-visualization.github.io/folium/



Preporučena literatura

- Edward Rolf Tufte (2001.), The Visual Display of Quantitative Information, 2nd ed., Graphics Press
- Cole Nussbaumer Knaflic (2019.), Storytelling with Data: Let's Practice!, 1st ed., Wiley

Zaključci

- Vizualizacija podataka omogućuje bolji uvid u podatke i rezultate te poboljšava komunikaciju sudionika na projektu
- Postoji veliki izbor grafova, no potrebno je odabrati onaj prikladan za određenu svrhu
- Koristiti vizualizaciju za otkrivanje različitih neobičnosti u podacima
- Pri prikazu, obratiti pozornost na boje i oblike tako da se prenese maksimalna informacija na što jednostavniji način
- U Pythonu postoji veliki broj alata za vizualizaciju, besplatnih i komercijalnih, većinom dobre kvalitete