

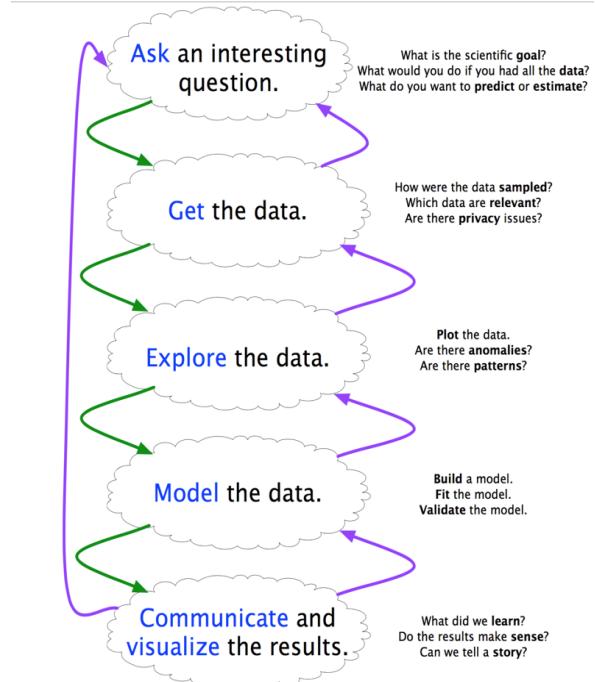
Sistemi za istraživanje i analizu podataka

Eksplorativna Analiza Podataka

predavač: Aleksandar Kovačević

Čemu služi?

- 1. Zajedno sa *data wrangling* čini prvi korak u procesu analize podataka u kome koristimo vizualizacije i sumarne statistike da obradimo: nedostajuće vrednosti, duplike i autlajere.
- 2. EDA takođe služi i za:
 - Otkrivanje novih šablonu u podacima (Vaš mozak je ovde primarni alat, vizualizacija je sekundarni).
 - Pričanje priče pomoću podataka (*data storytelling*).
- Fokus ovog predavanja je tački 2. sa ciljem da:
 - Dobijete savete o grafikonima.
 - Naučite kroz primere kako se kreira priča pomoću podataka.



Neobičan uvodni primer

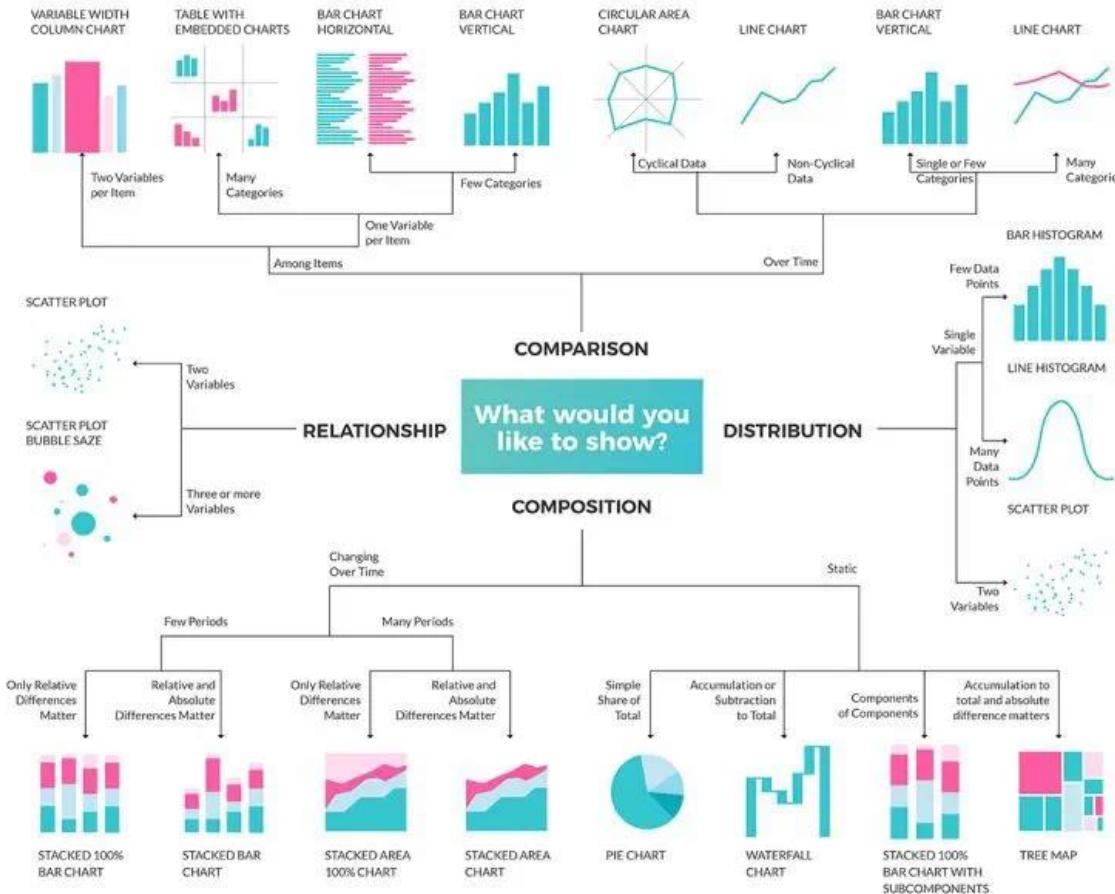


["Garden of Eden"](#): 8 glavica salate, od kojih je svaka smeštena u posebnu hermetički zatvorenu pleksiglas kutiju i predstavlja veliki grad. Koncentracija ozona u svakoj kutiji se u realnom vremenu kontroliše kako bi odražavala trenutni nivo zagađenja u gradu.

Deo 1

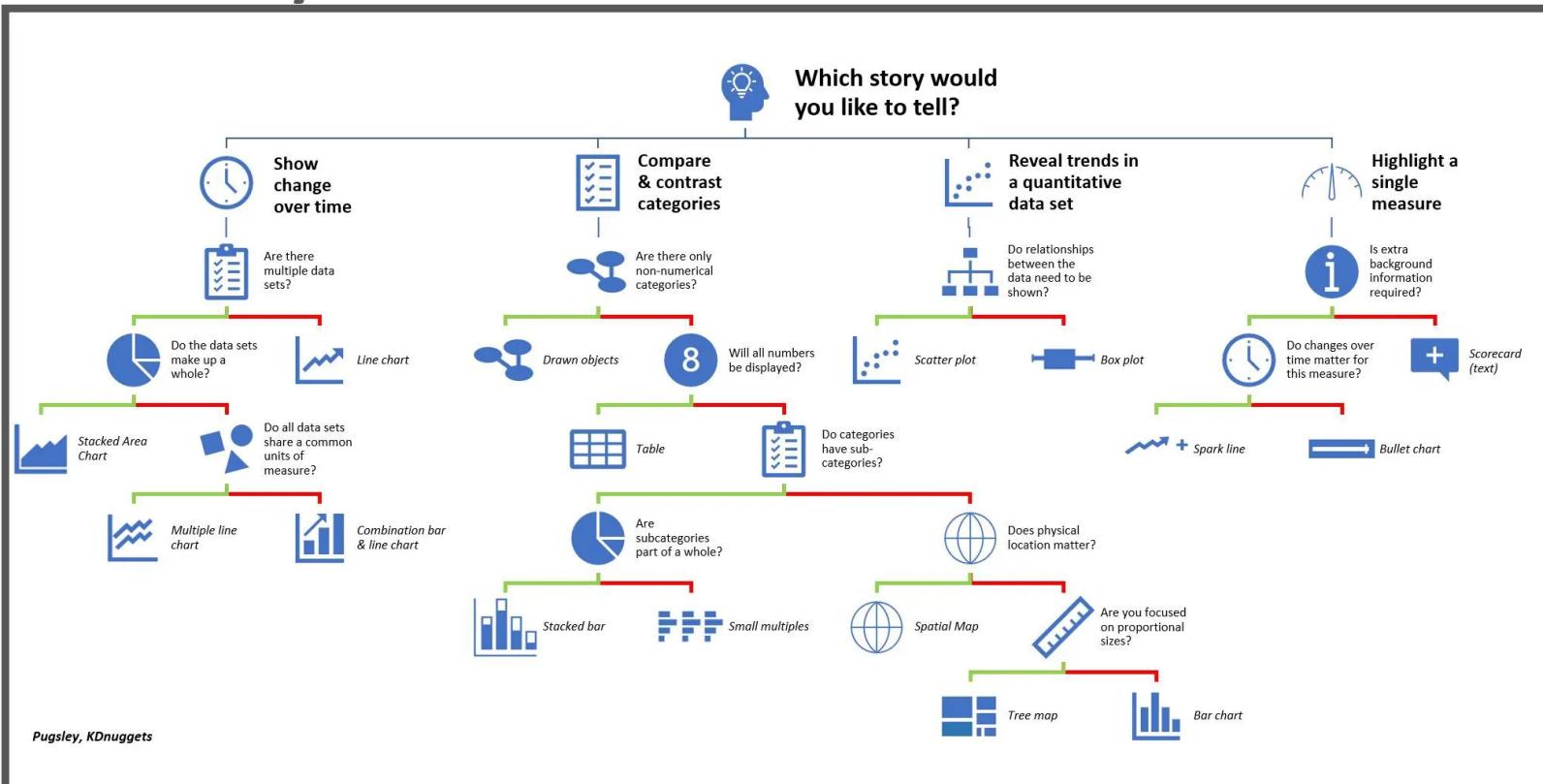
Kako odabratи odgovarajuћи grafik

Grafik sa savetima 1/2



Grafik sa savetima 2/2

Data Story Visualization: A Decision Tree



Pugsley, KDnuggets

Tabla sa savetima

You need to:

You should use:

Show data trends over time

Line

Compare categories

Bar, Pie

Compare totals

Pie, Stacked Bar

Show relationships

Scatter, Line, Facet, Dual Line

Show distributions

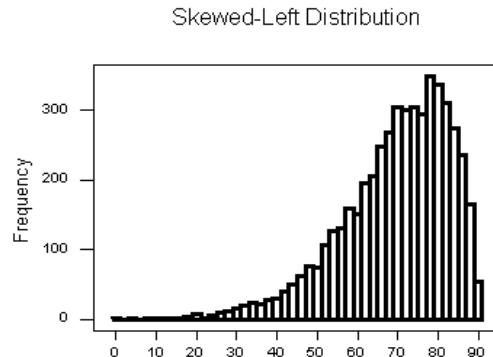
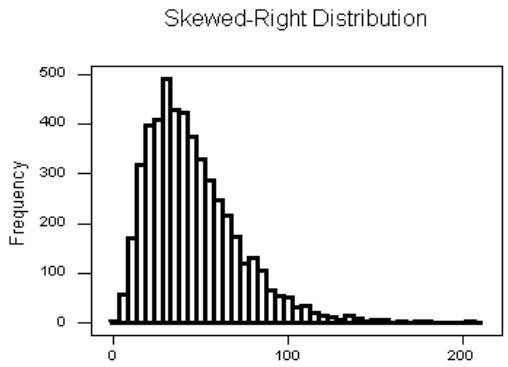
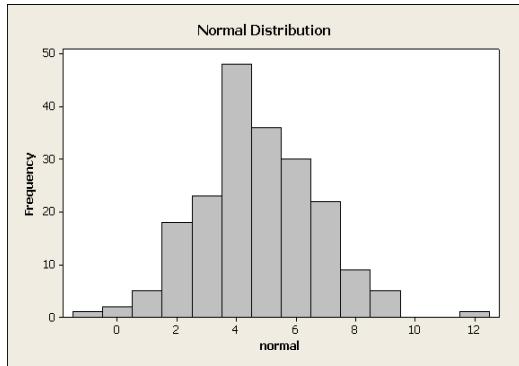
Scatter, Histogram, Box

Show proportions

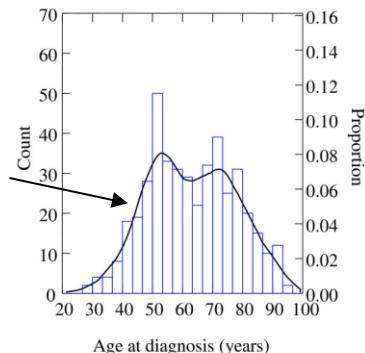
Pie, Donut, Waffle

Jedan atribut: histogrami

Mogu da otkriju mnogo toga o jednom atributu (diskretnom ili kontinualnom)

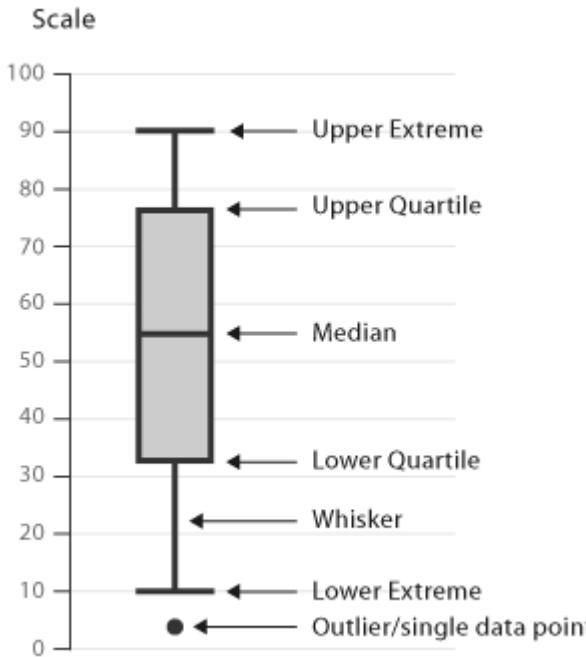


Izravnati (*smoothed*)
histogram
(prikaz distribucije
atributa)

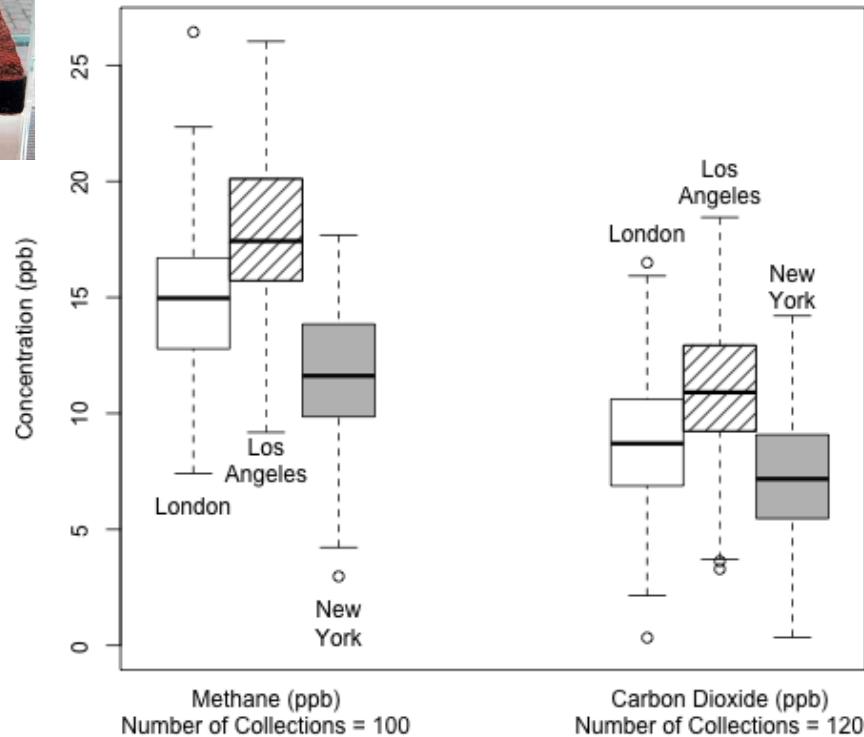


Olakšavaju
prepoznavanje
iskrivljenih
distribucija!

Jedan atribut: kutijasti dijagrami (*box plots*)



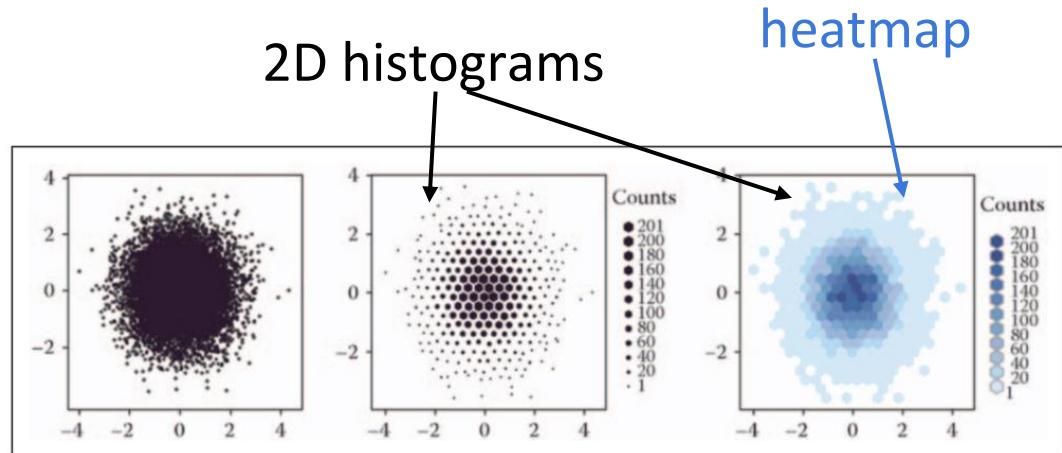
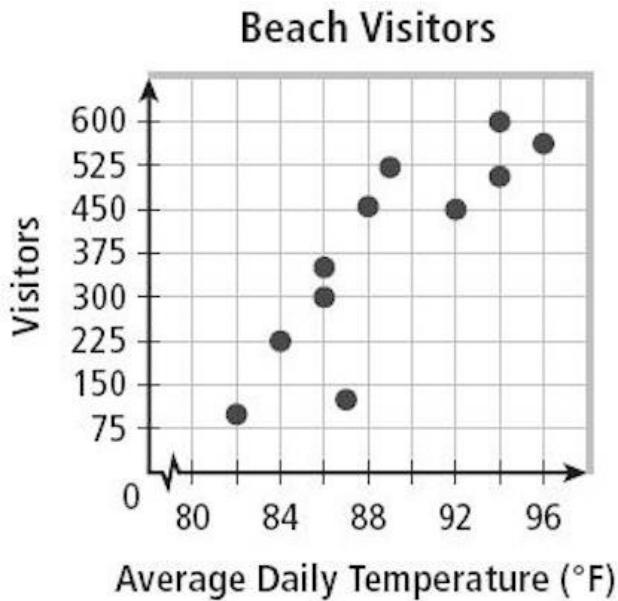
Comparing Pollution in London, Los Angeles, and New York



Dva atributa: dijagrami rasipanja (*scatter plots*)

Lako se mogu videti odnosi između dva atributa

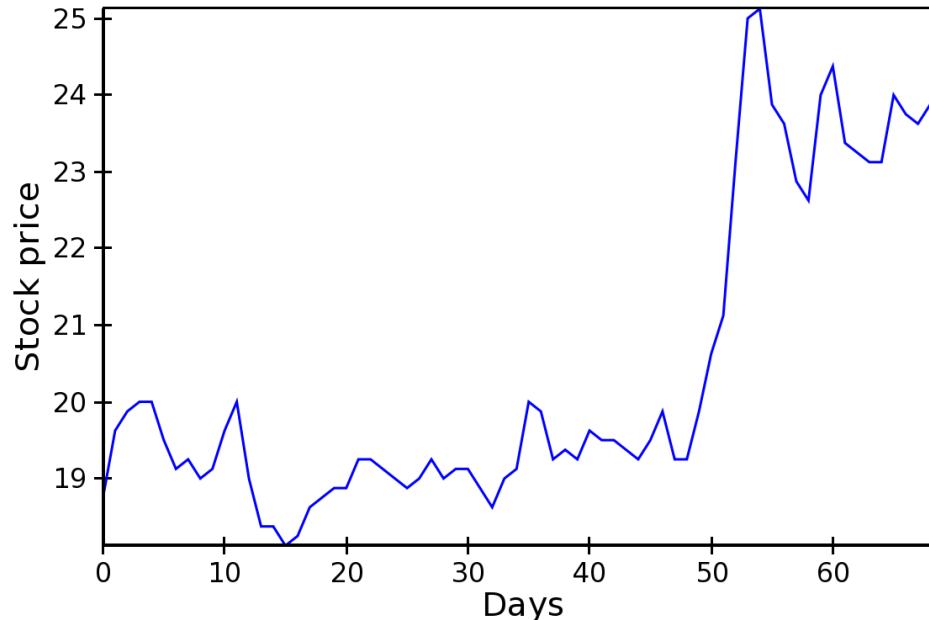
U slučaju velike količine podataka koristimo 2D histograme (toplote mape)



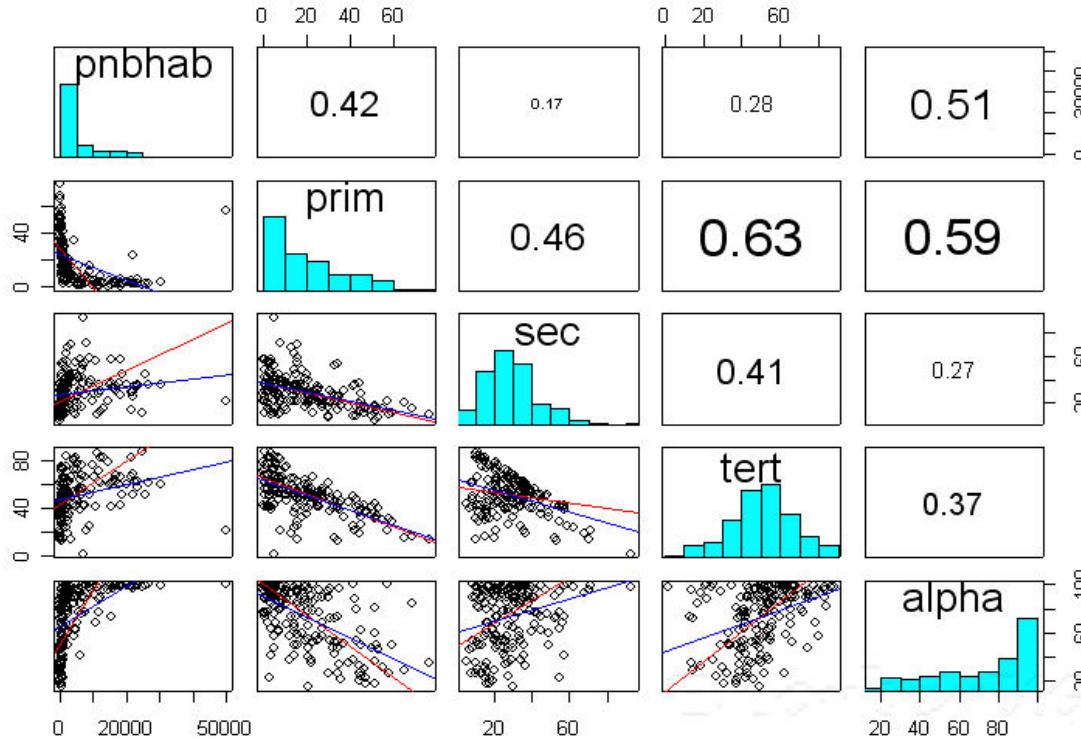
Dva atributa: dijagrami sa linijama (*line plots*)

Pogodni kada je odnos dva atributa neka funkcija (može biti i nepoznata)

Vrlo često se koriste kada je atribut na x-osi **vreme**.



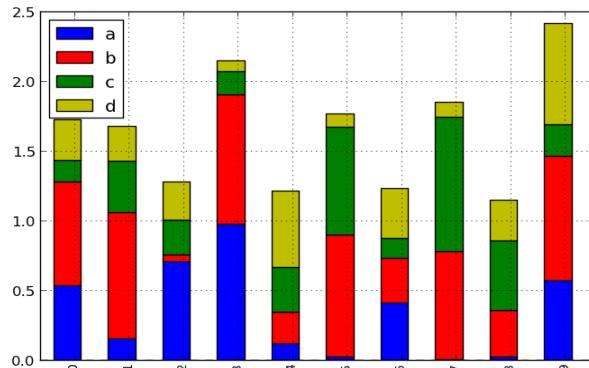
> 2 atributa: matrica dijagrama rasipanja (*scatter plot matrix*)



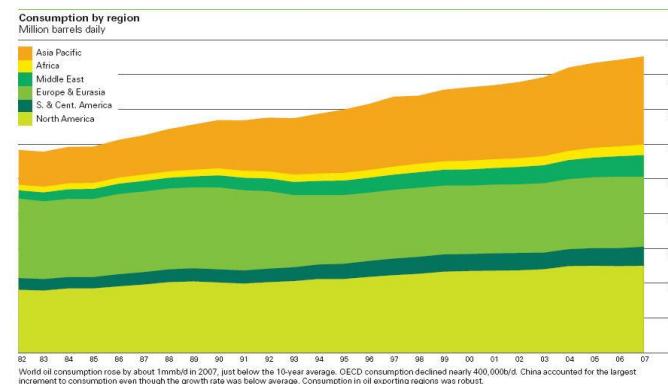
> 2 atributa: „naslagani“ dijagrami (*stacked plots*)

Ovde: 3 atributa: pozicija, visina (debljina) i boja

Atribut koji je naslagen i za koji se koristi boja je kategorički, visina predstavlja kontinualni atribut.

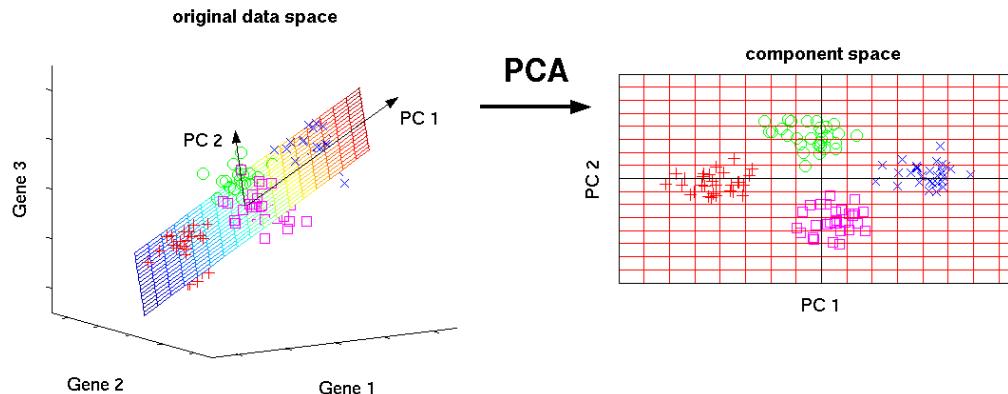


Boja je kategorička, druga dva atributa su kontinualni.

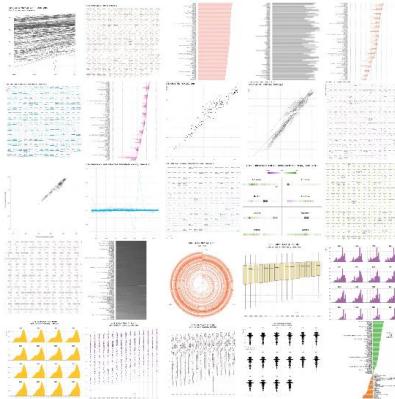


Redukcija dimenzionalnosti

- Mapiranje visoko-dimenzionog prostora u 2d ili 3d.
- PCA je tipičan primer
 - Prednost: koordinatne ose su interpretabilne
 - Mana: podrazumeva linearni odnos atributa u visoko-dimenzionom prostoru
- Postoji mnogo alternativa: Pacmap, Umap i T-sne su najpoznatije.



Jedan skup podataka vizualizvoan na 25 načina



<http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways>

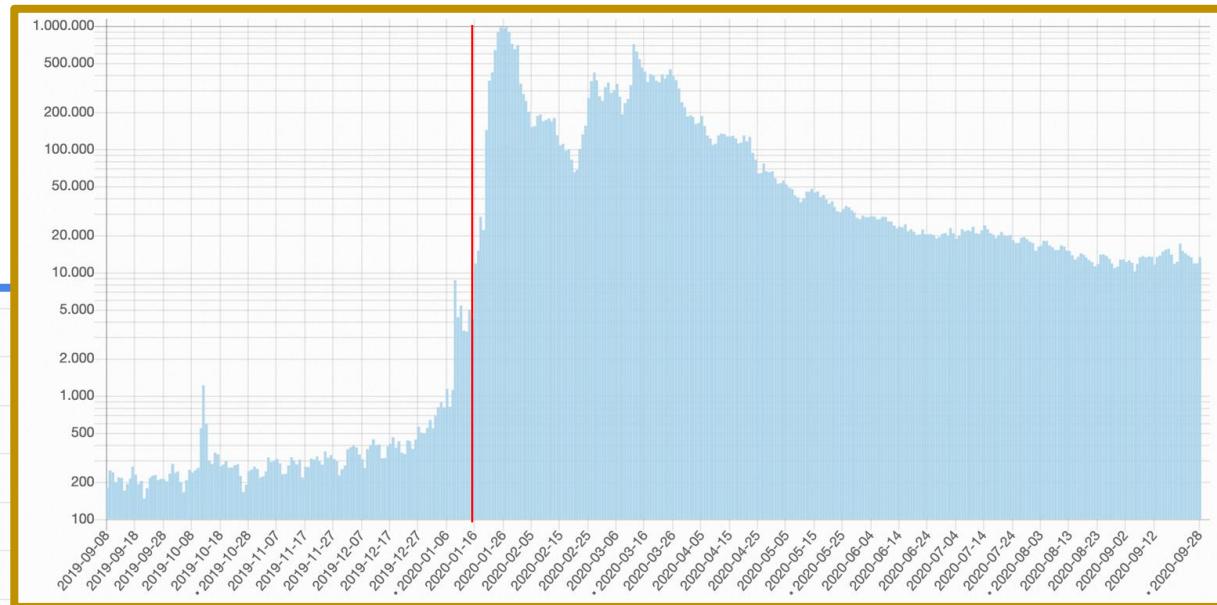
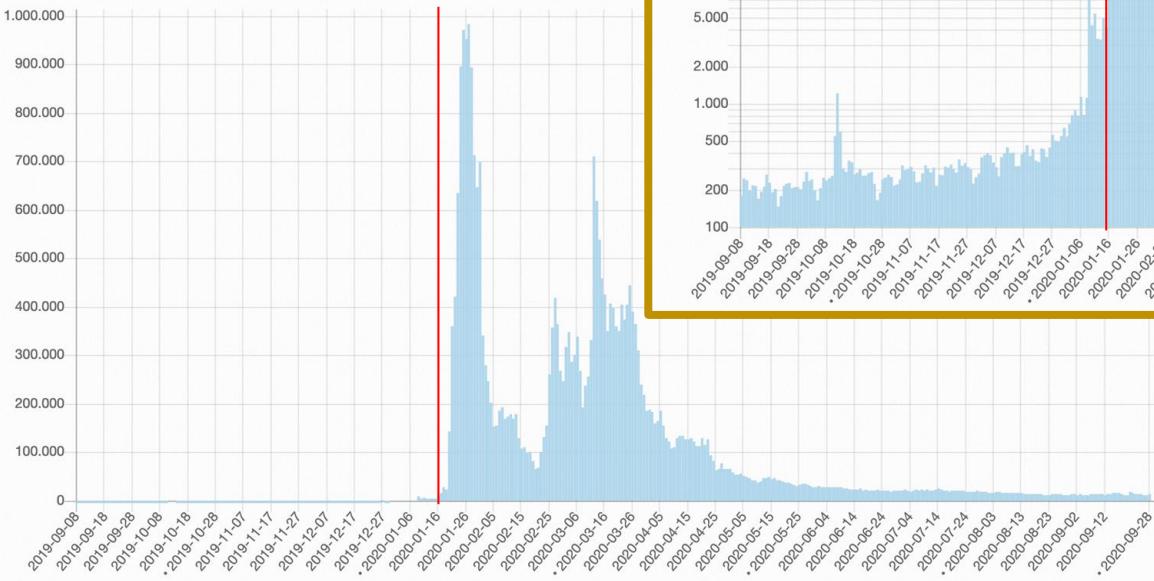
“You must help the data focus and get to the point. Otherwise, it just ends up rambling about what it had for breakfast this morning and how the coffee wasn’t hot enough.”

Deo 2

Dodatni saveti oko kreiranja grafikona

Odabir koordinatnih osa je važan!

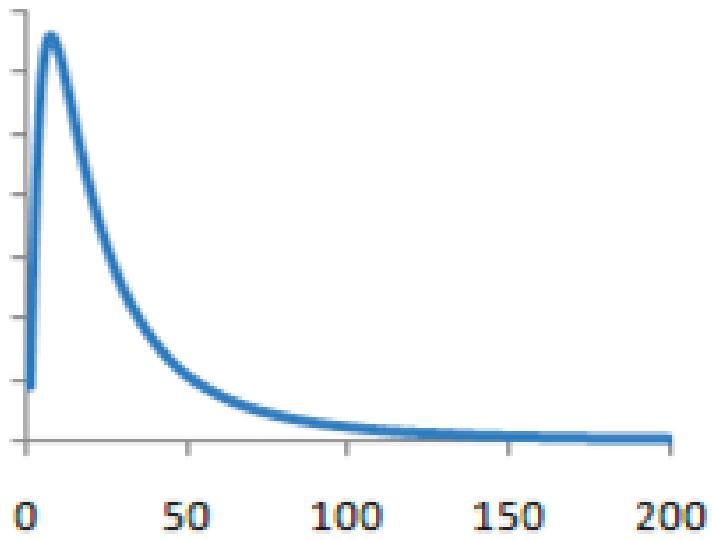
Vremenska serija sa
brojem pregleda Wikipedia
stranice: “Coronavirus”
(linearna y-osa)



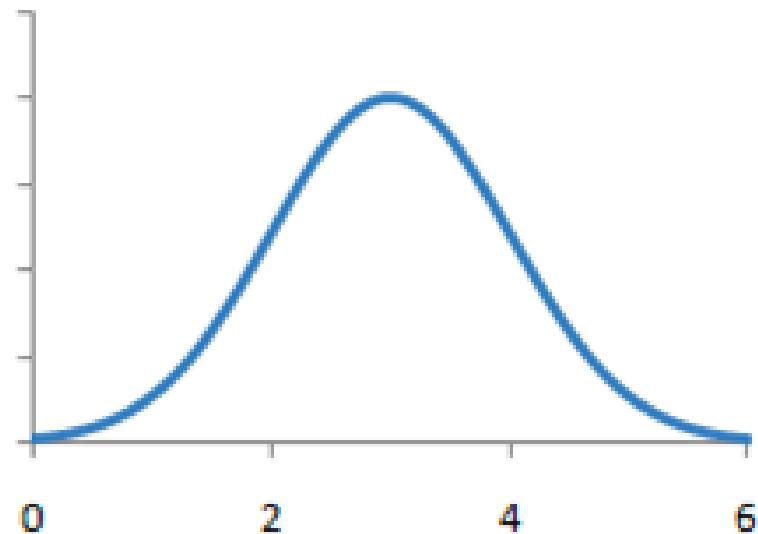
(logaritmovana y-osa)

Odabir koordinatnih osa je važan:

Prikaz distribucija koje imaju izražen rep



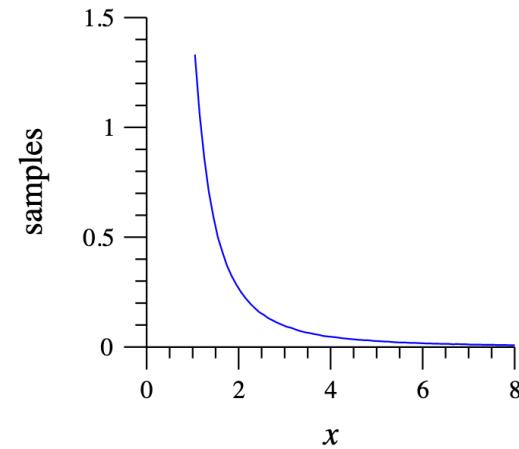
Linearna x-osa



Logaritnomvana x-osa

Distribucije stepenog zakona (*Power laws*)

- Matematički oblik, verovatnoća x : $p(x) = Cx^{-\alpha}$,
- Često se nazivaju i pravilo 80/20 ili Paretove distribucije
- Grubo gledano 20% x vrednosti čine 80% vrednosti
- Mnogi fenomeni u društvu su power laws:
 - Populacija u gradovima i selima
 - 80% populacije živi u 20% nastanjenih mesta
 - Računarstvo: popravka 20% najvećih bagova rešava 80% problema koji korisnici imaju
 - 80% zarade od 20% mušterija
 - 80% značajnih prijateljstava od 20% prijatelja...

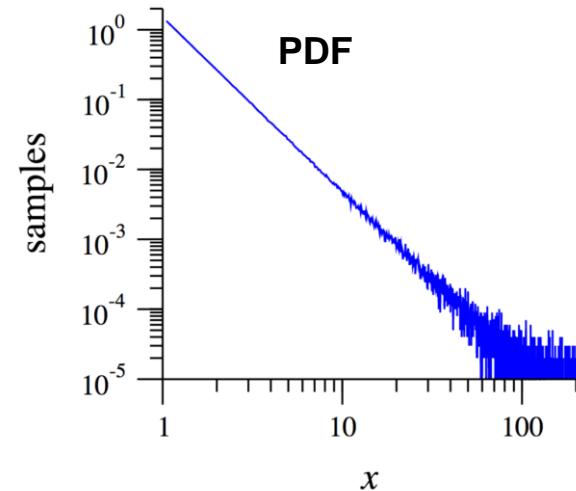
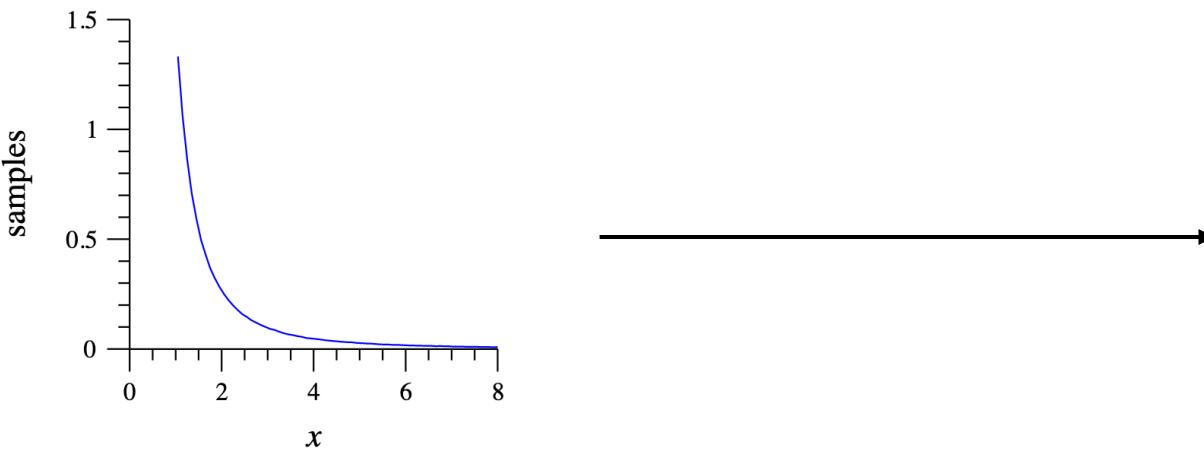


Power laws – kako ih prepoznati

- Daju pravu liniju na log-log dijagramu:

$$y = C x^{-\alpha} \Leftrightarrow \log(y) = \log(C) - \alpha \log(x)$$

- Od sumarnih statistika medijan je korisniji od srednje vrednosti.

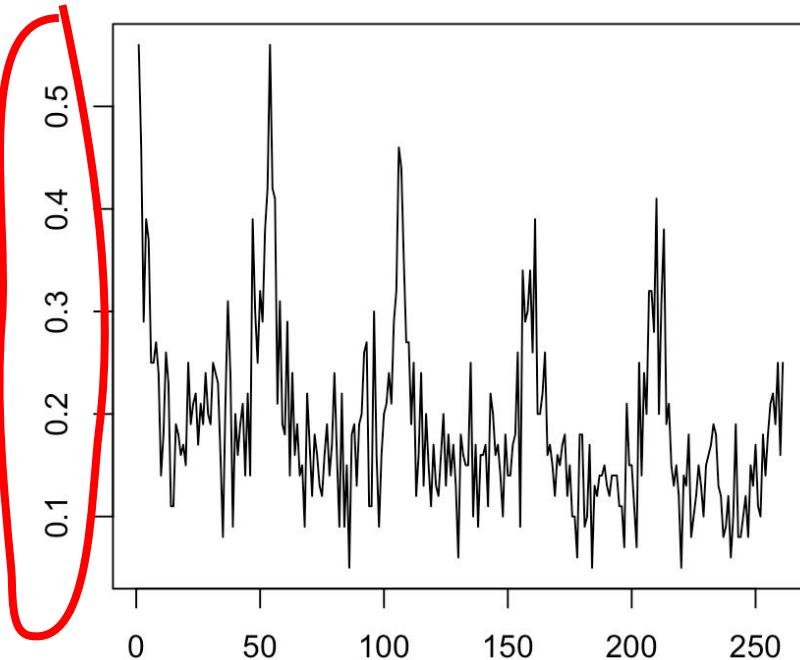
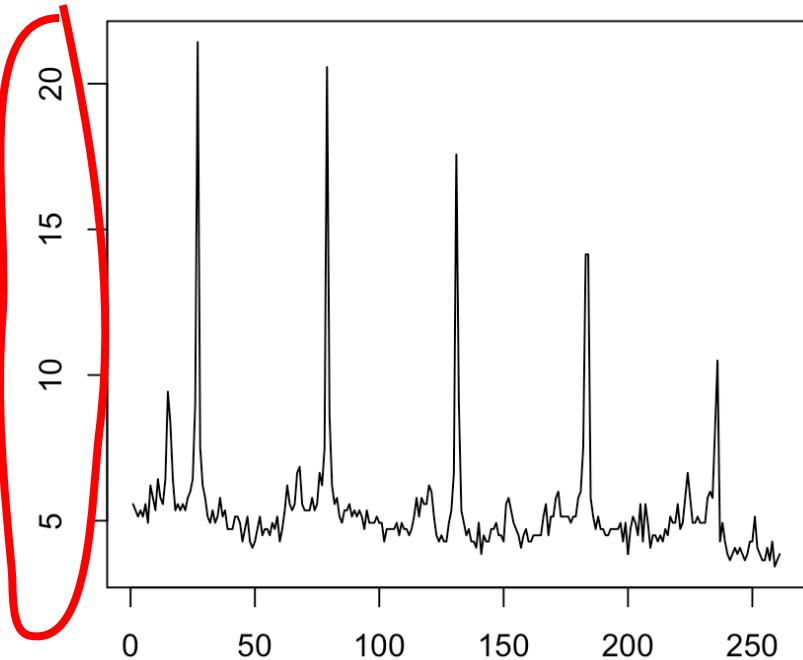


Power laws – zašto su važni

- Otkrivanje power laws može:
- Povećati produktivnost i zaradu:
 - Na primer, power law o bagovima ili zaradi i mušterijama...
- Dati nam uvid u procese funkcionisanja društva
 - Na primer, power law o populaciji, bogatstvu, prijateljstvu...
- Dati nam uvid u procese funkcionisanja prirode
 - Na primer, 80% živih organizama čini 20% vrsta, 80% od svih sinapsi u mozgu vezano je za 20% neurona...

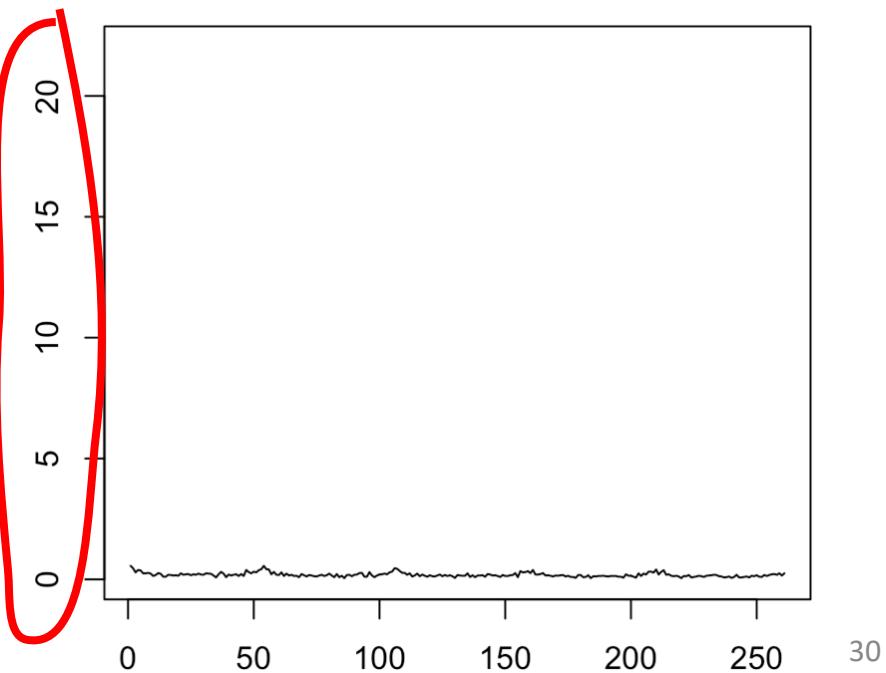
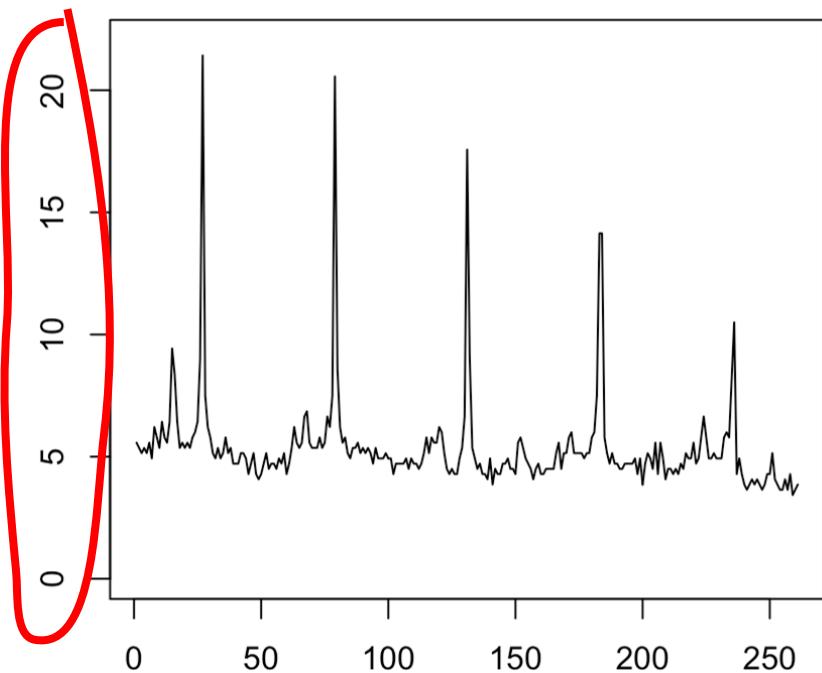
Koordinatne ose bi trebalo da budu dosledne!

Na prvi pogled: koja vremenska serija ima veću srednju vrednost?

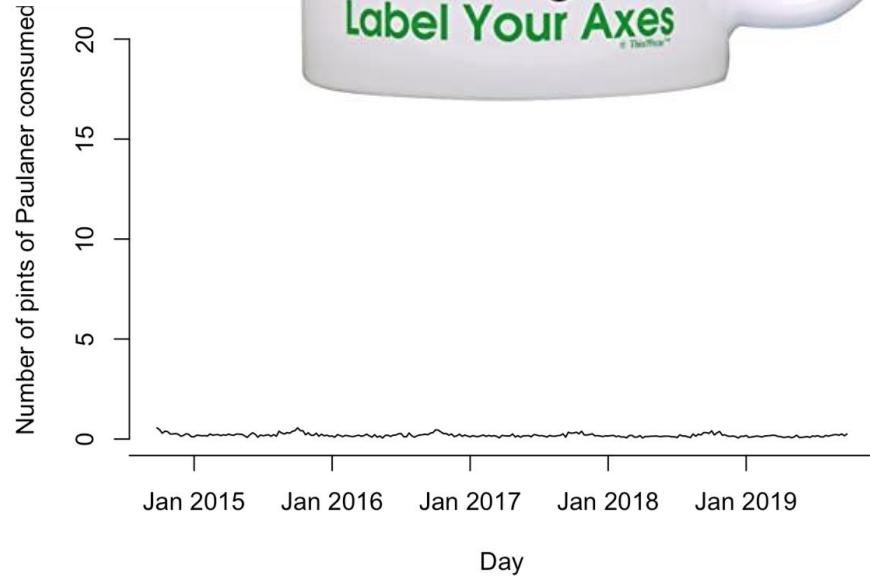
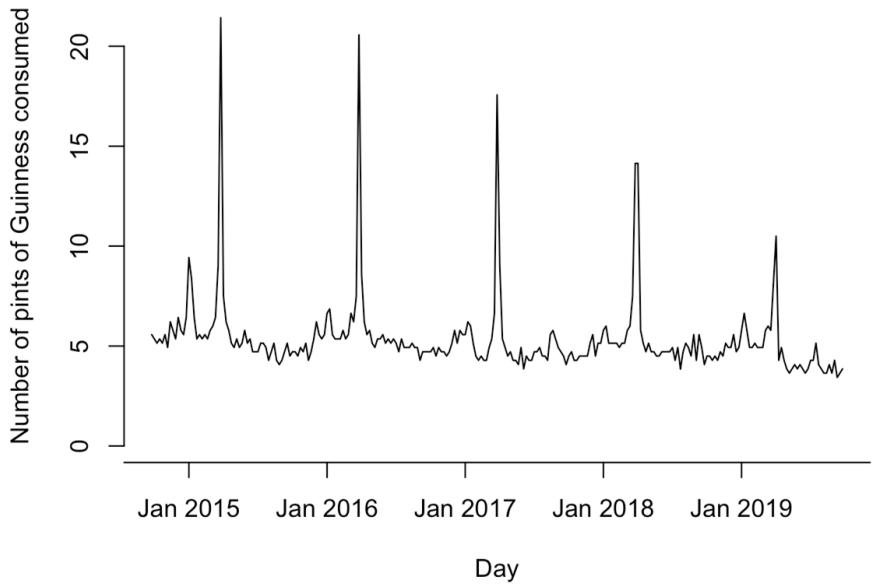


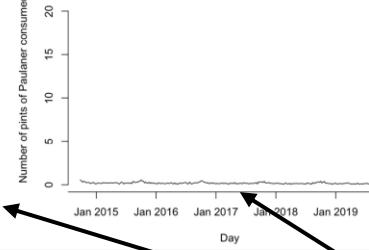
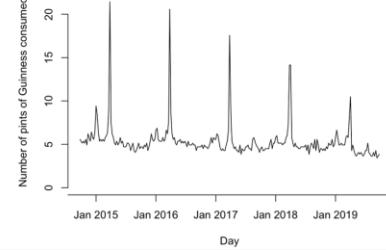
Koordinatne ose po bi trebalo da budu dosledne!

Na prvi pogled: koja vremenska serija ima veću srednju vrednost?

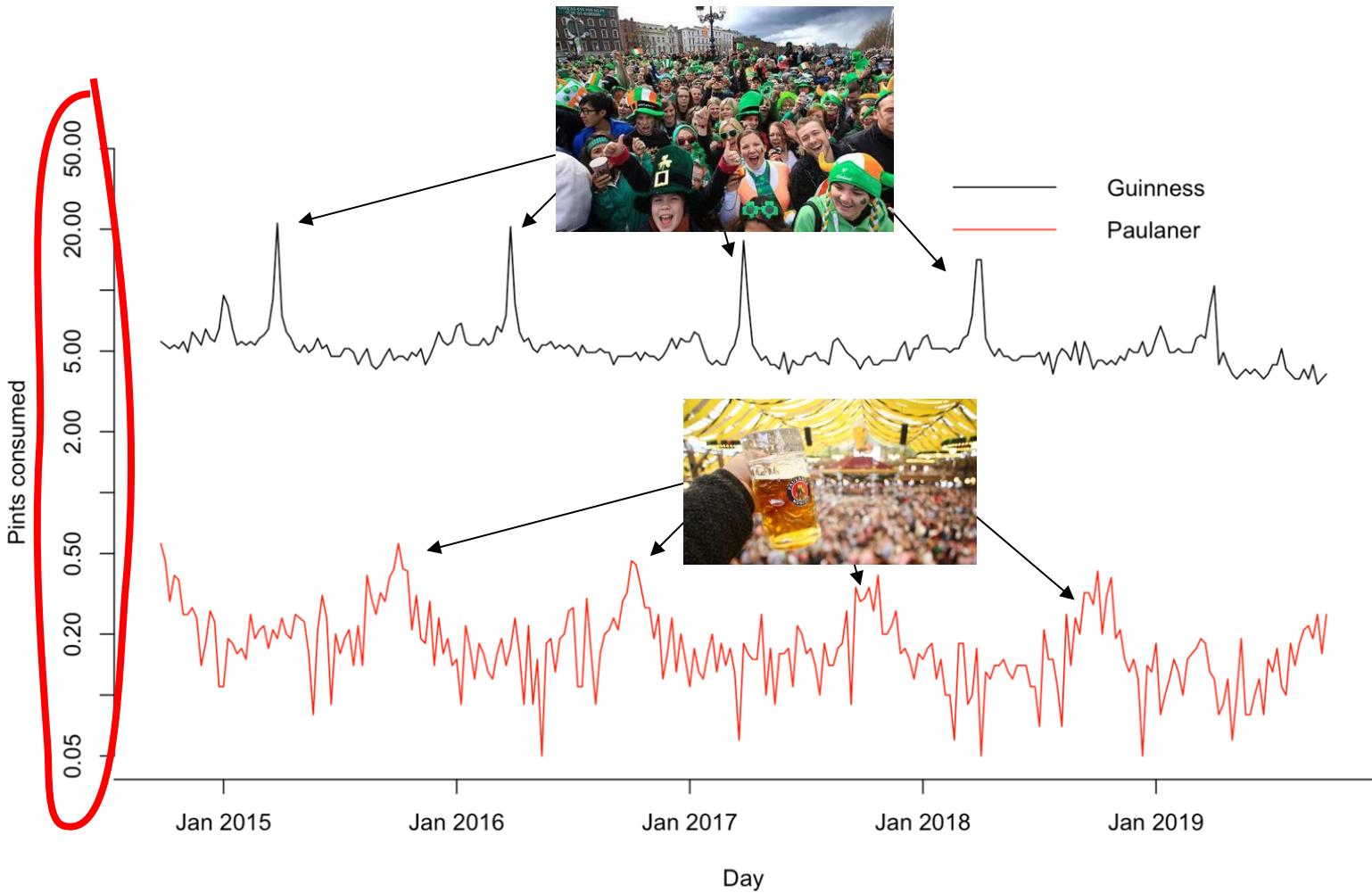


Oznake osa su važne!





Da li možemo da ih
spojimo u jedan grafik?



Izbor boja je važan!

Savet je da izaberete paletu u skladu sa onim što želite da prikažete:

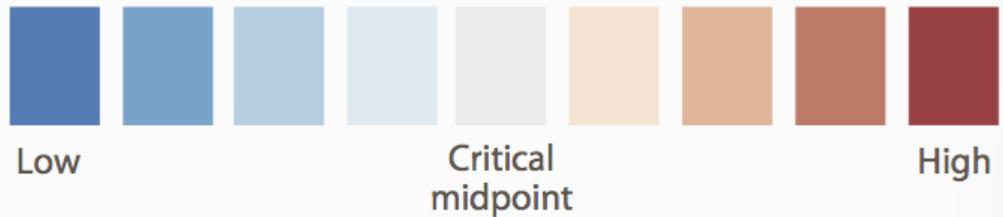
Sequential

Colors can be ordered from low to high



Diverging

Two sequential schemes extended out from a critical midpoint value

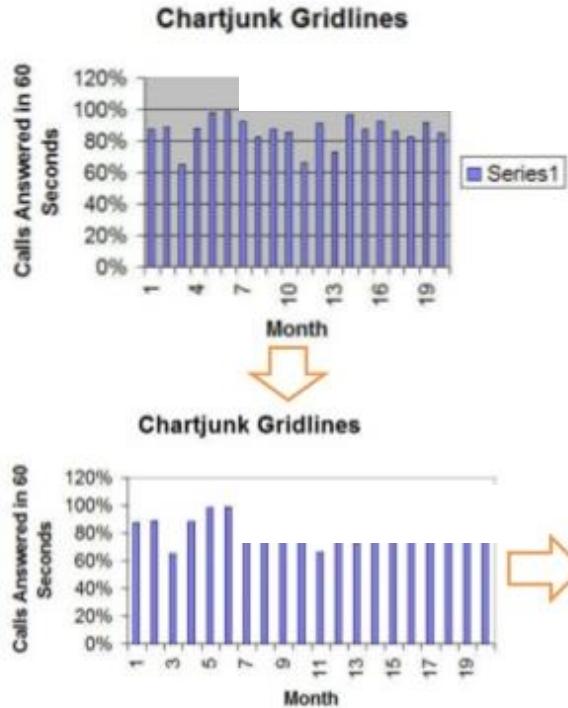


Categorical

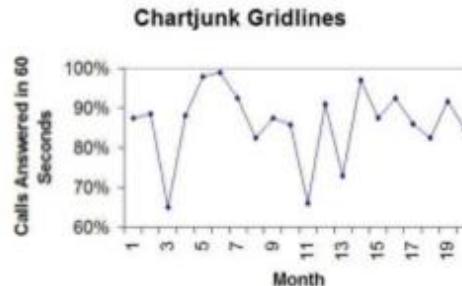
Lots of contrast between each adjacent color



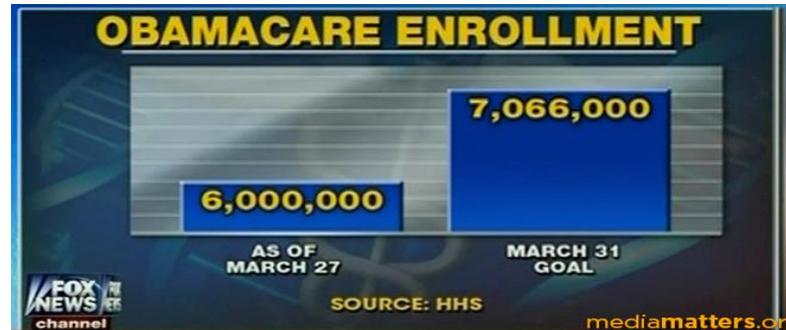
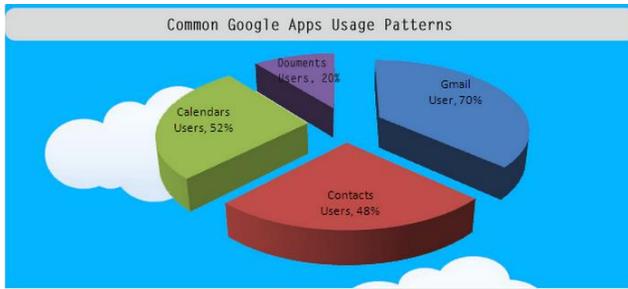
Ne pretrpavajte grafikone!



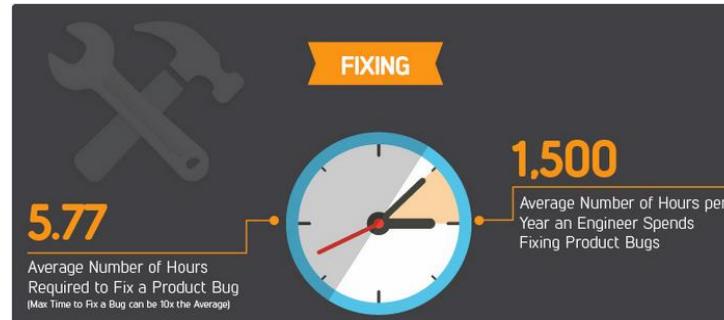
Graphical excellence gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
-Edward Tufte



Primeri loših grafikona

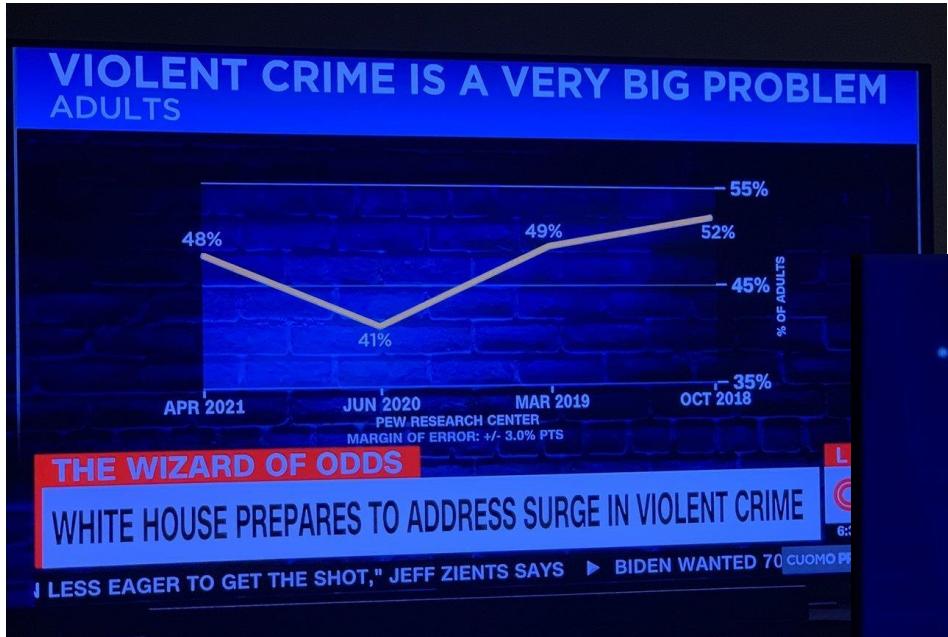


90% of US Households Consume Peanut Butter

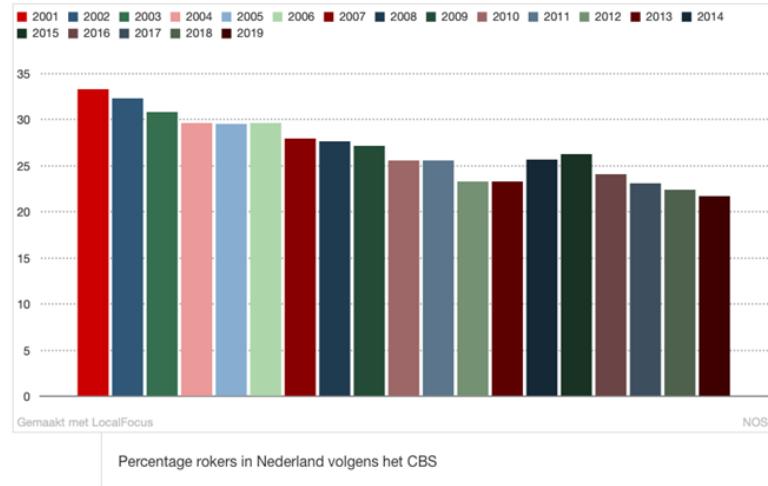
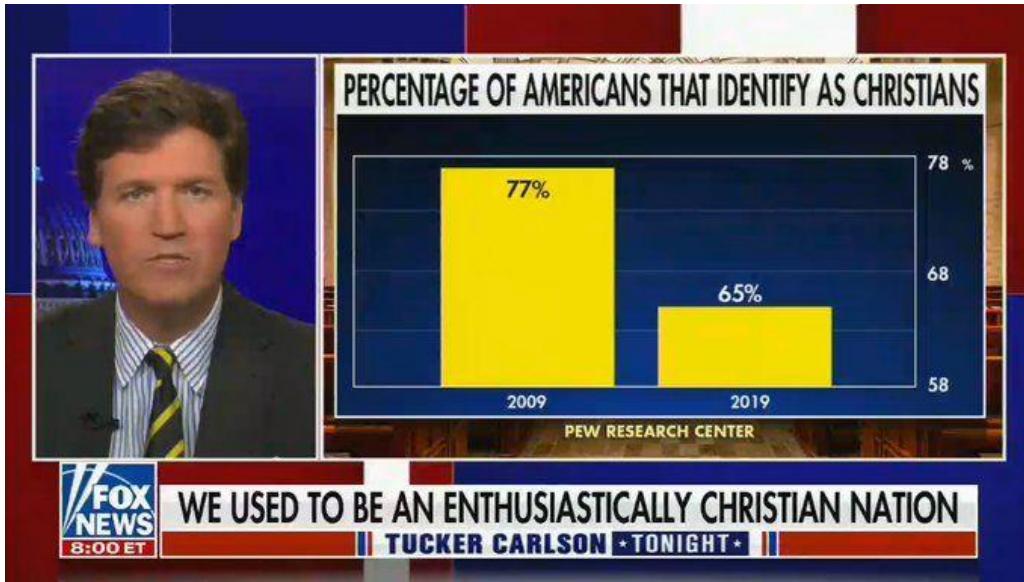


Courtesy of viz.wtf

Primeri loših grafikona

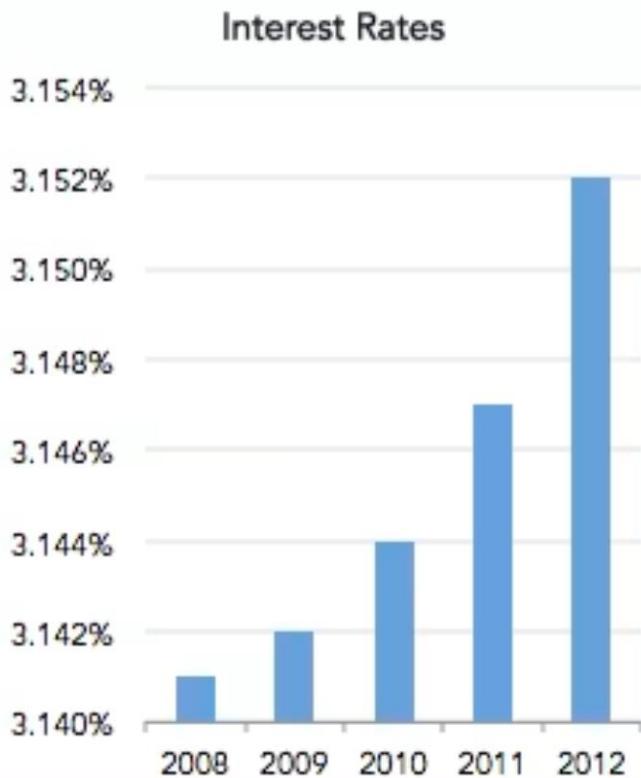


Primeri loših grafikona

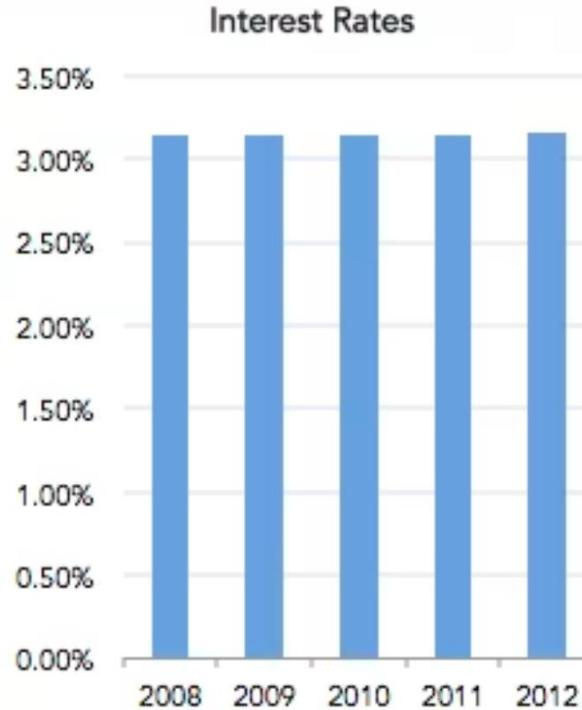


Primeri loših grafikona

Loše



Dobro



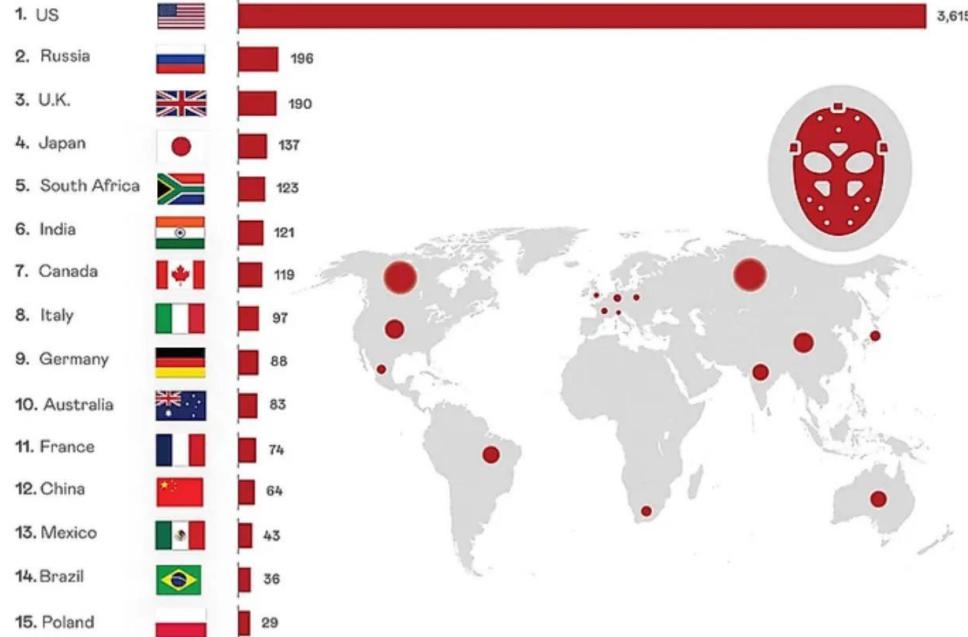
Primeri loših grafikona

© WorldAtlas Graphics



Serial Killers by Country

The US has produced the most known serial killers (3,615) followed distantly by Russia (196.)

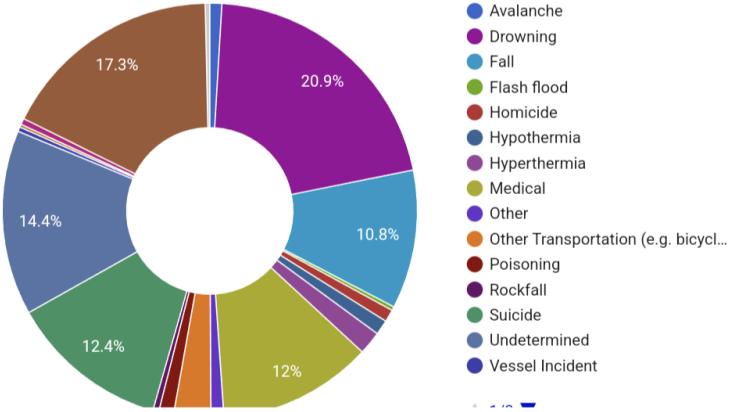


Primeri loših grafikona

Loše

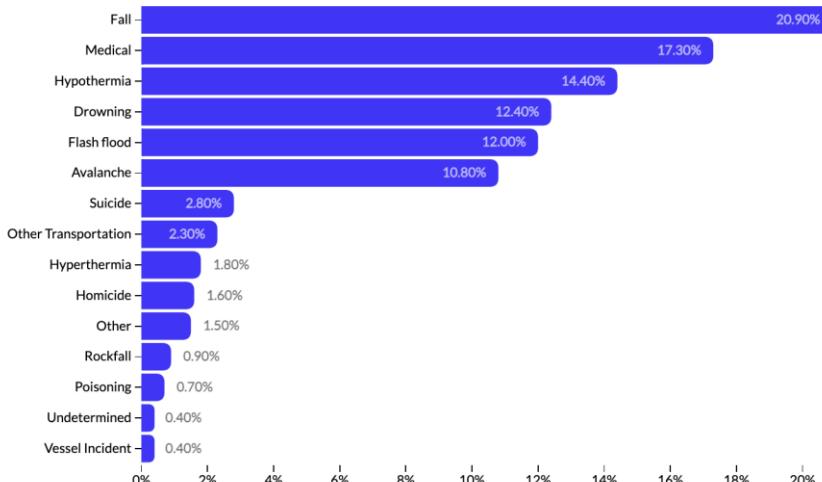
What Are the Top Causes of Death in the National Parks?

Fatalities in National Parks (2007-2023) by Cause of Death



Dobro

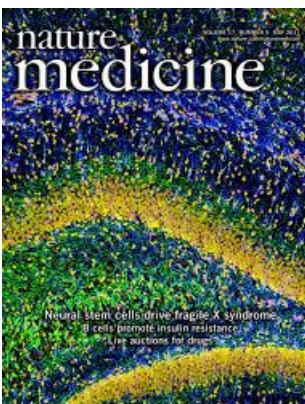
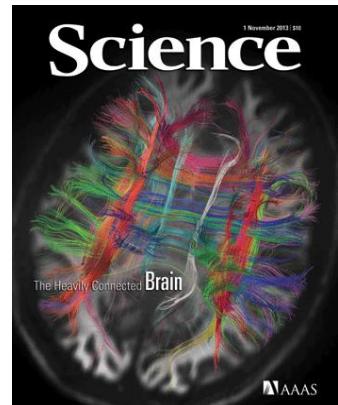
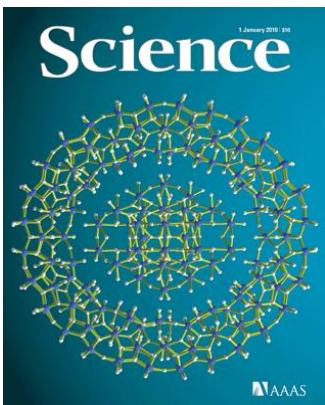
Top Causes Of Death In National Parks



Deo 3

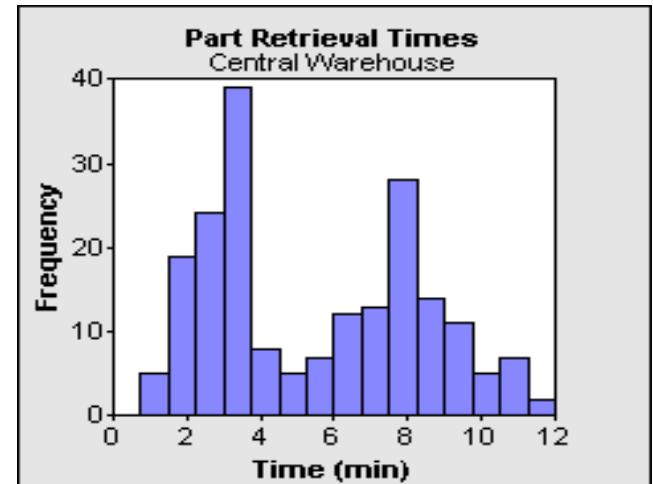
Slučajevi korišćenja (*use cases*) EDA

Prezentovanje naučnih rezultata



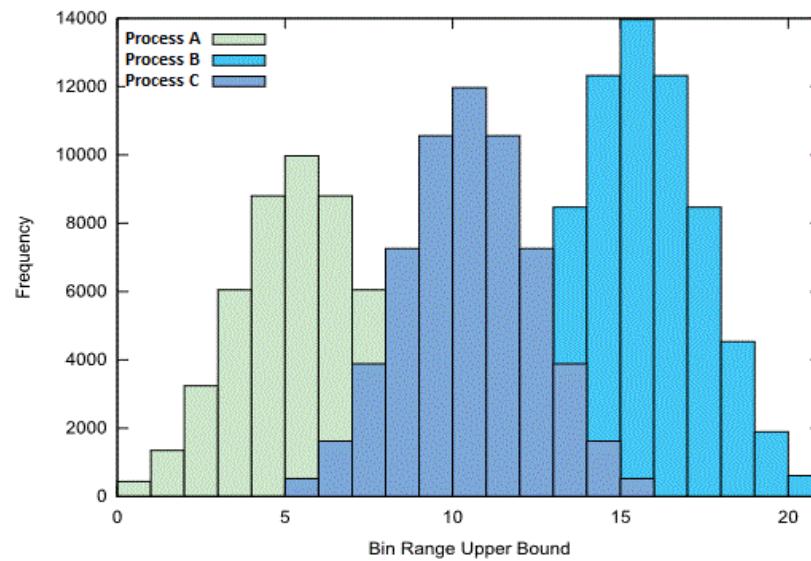
Podaci iz više distribucija

- Dva ili više različitih vrhova u histogramu često sugeriju 2 ili više različitih populacija čije uzorke imamo u podacima.
 - Dok mi npr. očekujemo samo jednu za taj skup podataka.
- Ali ne nagađajte! Istražite dalje koristeći, na primer, boju i histogram više populacija (sledeći slajd)



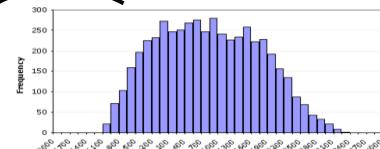
Podaci iz više distribucija

Istražite dalje koristeći, na primer, boju i histogram više populacija

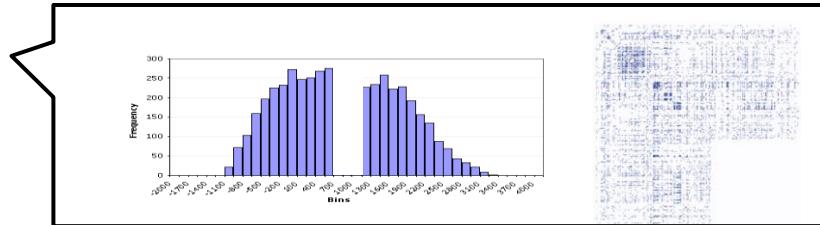


Data wrangling

Čudni podaci



- Održavajte teoriju o tome kako bi podaci trebali izgledati.
- Neki podaci su veoma teški za objašnjenje.
- Nikada ne ignorišite čudne podatke!
- Prvo, prepostavite da je u pitanju greška.
- Pokušajte da je ispravite.
- Ako nije greška: možda ste napravili zanimljivo otkriće!
- Neki od najvažnijih naučnih otkrića dobijeni su ne ignorisanjem čudnih podataka, već njihovim analiziranjem (npr. *cosmic background radiation*).



Novinarstvo (data storytelling)

NY Times interactive visualizations (recession/recovery 2014)

<http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html>

And 2014 “the year in interactive storytelling”

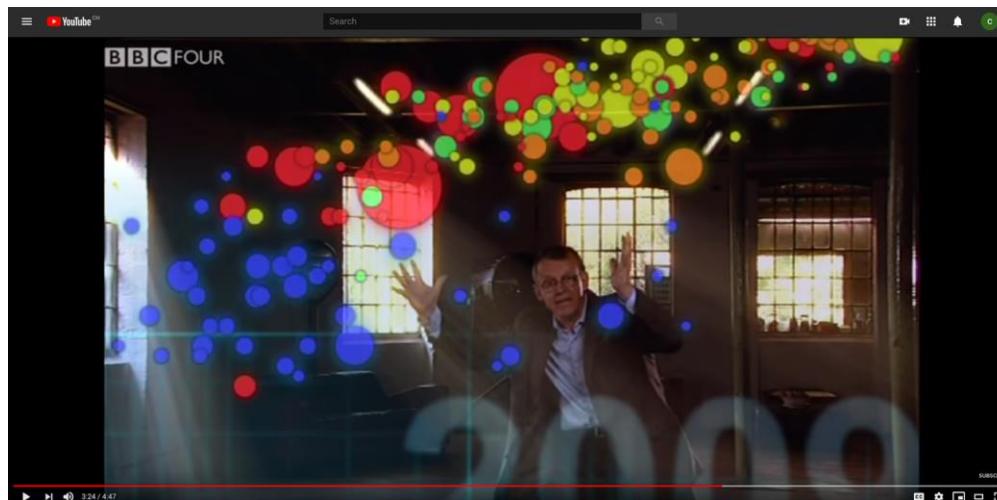
http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html?_r=0

Obrazovanje šire javnosti

Hans Rosling:

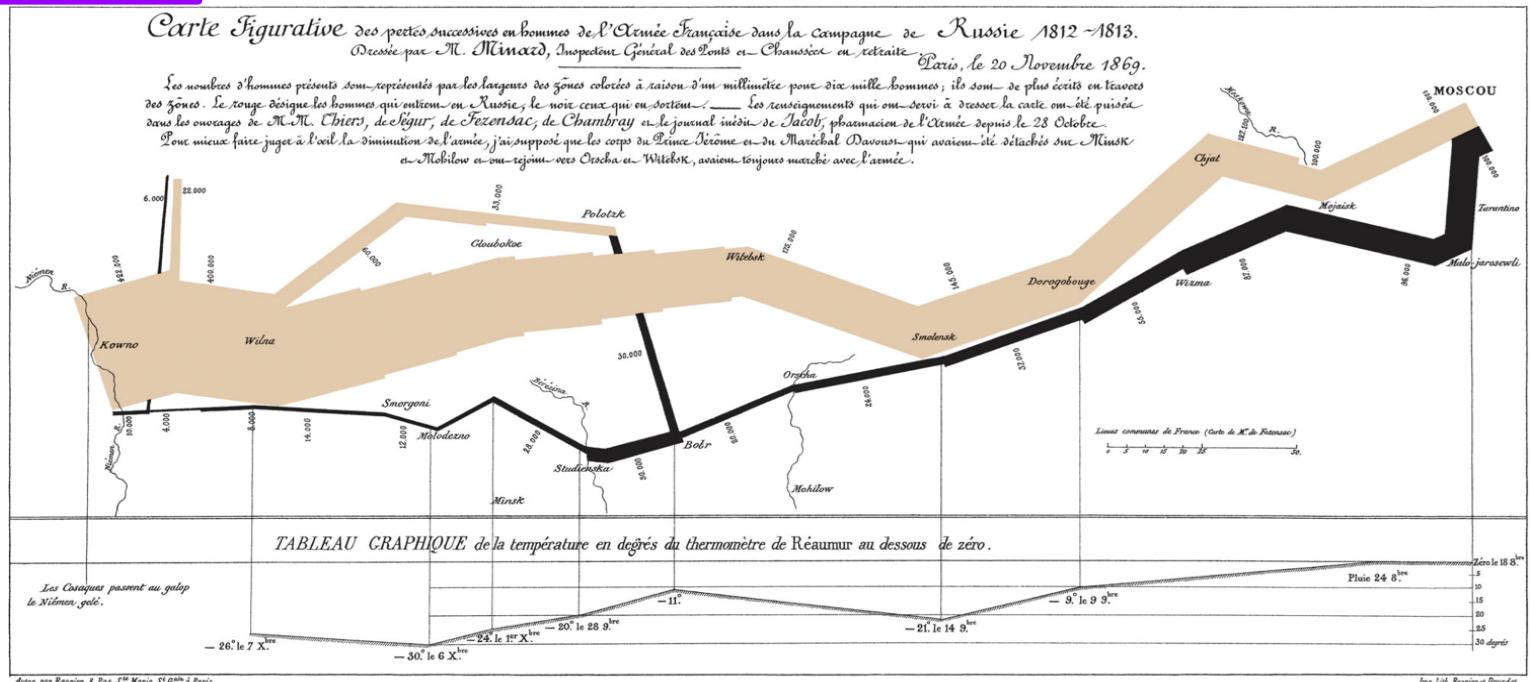
200 countries, 200 years, 4 minutes

<https://www.youtube.com/watch?v=jbkSRLYSOjo>



Pružanje novih uvida u podatke

Charles Joseph Minard 1869 Napolenov marš na Rusiju



"It may well be the best statistical graphic ever drawn."

5 atributa: veličina vojske, lokacije, datumi, pravci kretanja, tempreatura vazduha

Primeri Eksplorativne Analize Podataka

SIAP projekata prethodnih generacija

Napomena: prikazani su samo interesantni grafici, ne komplenti EDA procesi

„First booking prediction based on – Airbnb New User Bookings data set“

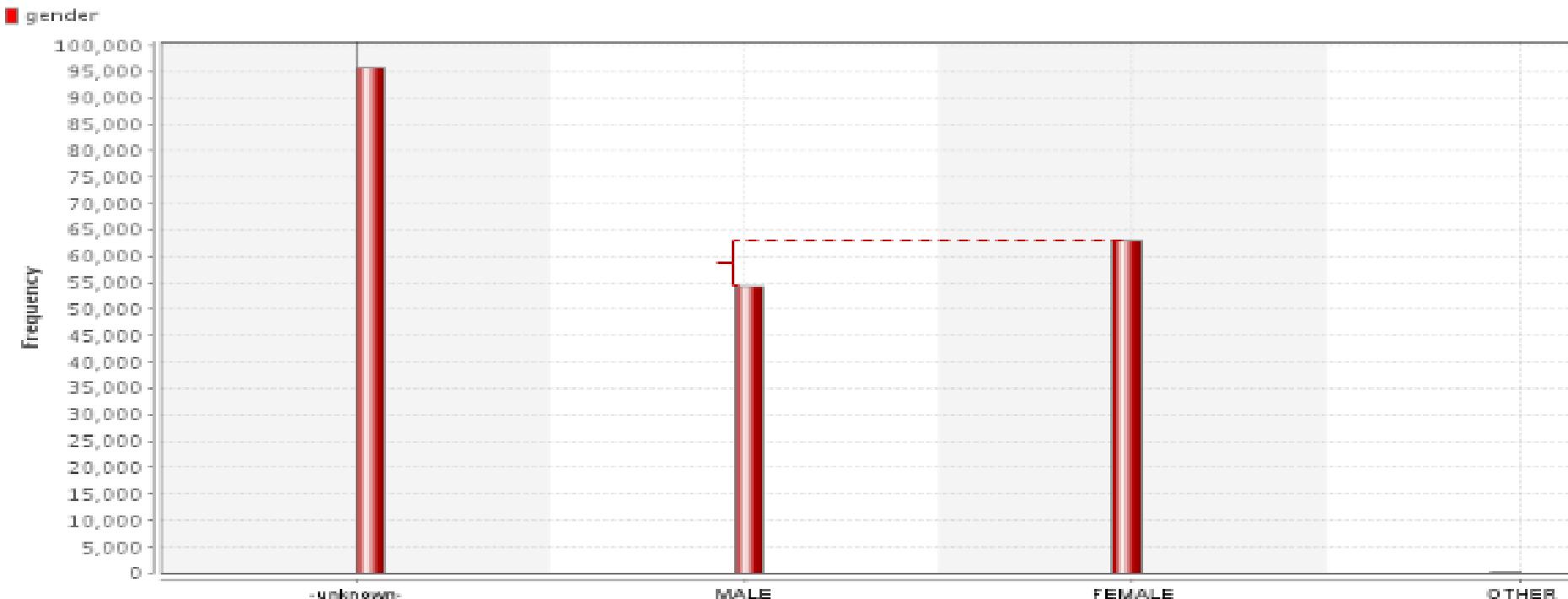
- Autori: Aleksandar Bošnjak, Ivana Živić, Violeta Novaković
- Cilj: na osnovu prethodne aktivnosti korisnika predvideti datum kada će izvšiti prvu rezervaciju
- Atributi:
 - Podaci o korisniku: id, date_account_created, timestamp_first_active, date_first_booking, gender, age, language....
 - Podaci o sesijama: user_id, action, action_type, action_detail, device_type, secs_elapsed.

„First booking prediction based on – Airbnb New User Bookings data set“



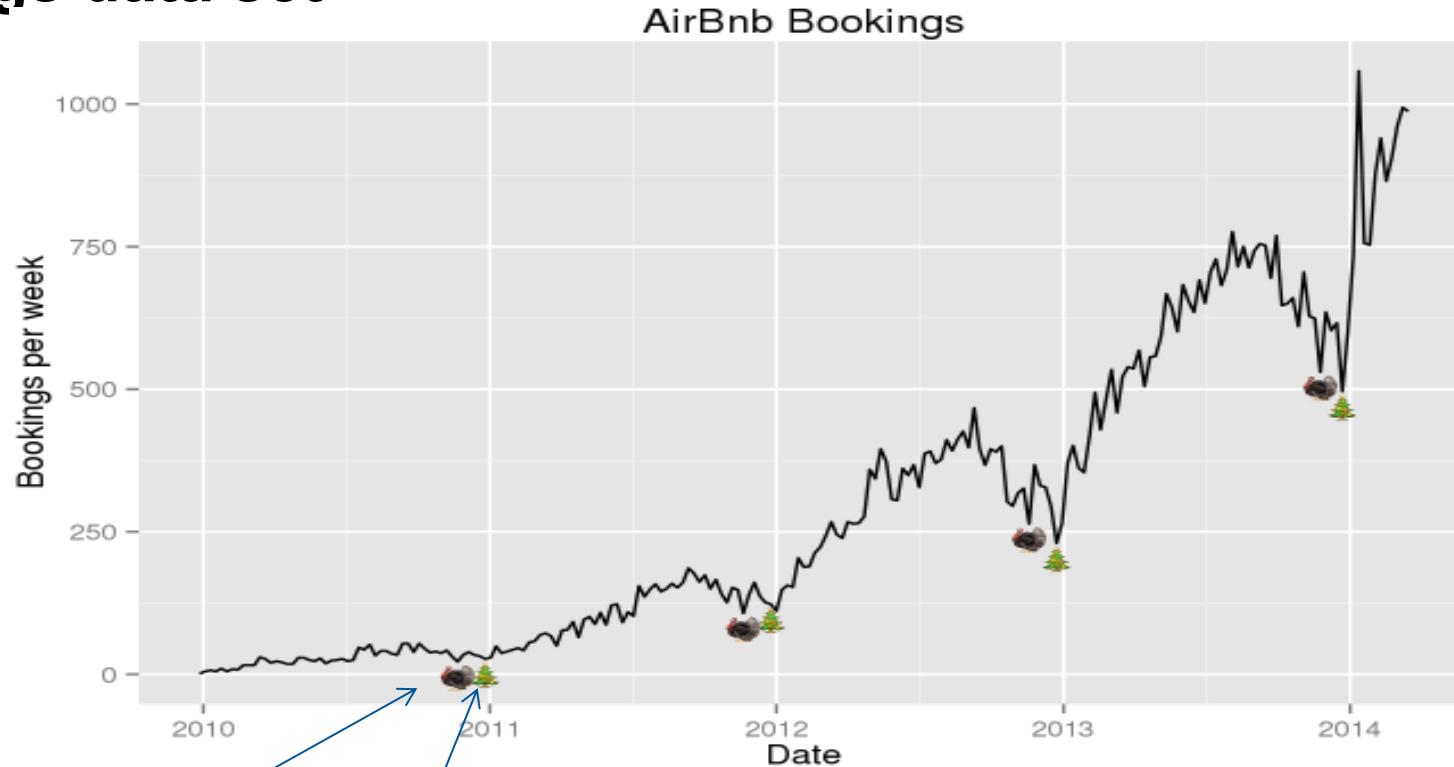
Age atribut. Autlajeri: 0 i 2014 (ocigledno los unos)

„First booking prediction based on – Airbnb New User Bookings data set“



Gender atribut. Dosta ljudi nije htelo da otkrije pol. Više korisnika ženskog pola.

„First booking prediction based on – Airbnb New User Bookings data set“



Novo i zanimljivo saznanje: Ljudi sve više izbegavaju da putuju praznicima (Dan Zahvalnosti i Božić)

„Analiza Avionskih Letova na Području Sjedinjenih Američkih Država, Analiza kašnjenja“

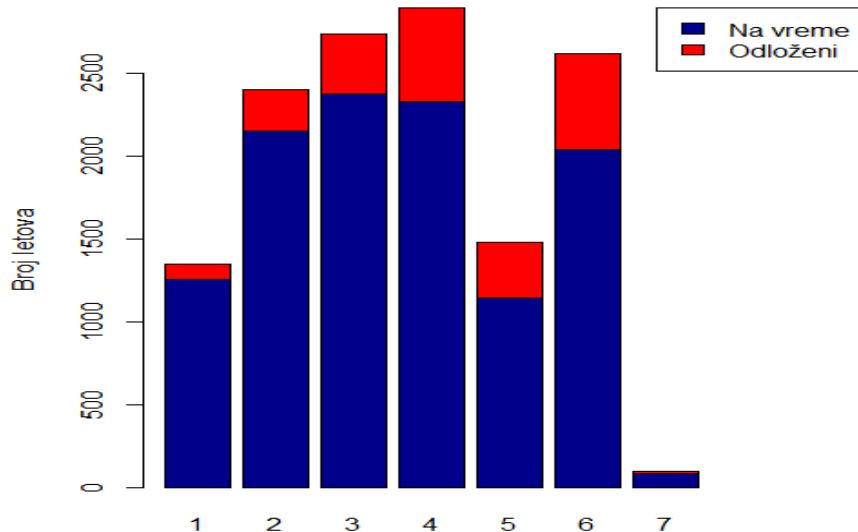
- Autor: Tamara Milovanović
- Cilj: pronalaženje glavnih uzroka velikog broja odloženih letova:
 - Za koji period dana je najmanja verovatnoća kašnjenja ili otkaza leta?
 - Šta su najčešći uzroci kašnjenja ili otkaza?
 - Koji tipovi aviona najčešće kasne?
 - Na kojim rutama je kašnjenje najčešće?
 - Koliki uticaj imaju vremenske nepogode na odlaganje letova?

„Analiza Avionskih Letova na Području Sjedinjenih Američkih Država, Analiza kašnjenja“

- Podaci:
- Podaci o avionima, letovima i aerodromima iz baze Američkog departmana za transport
- Informacije o avio prevoznicima, tipovi aviona itd. <http://www.airfleets.net/home/>
- Podaci o vremenskim uslovima preuzeti iz NOAA (National Oceanic and Atmospheric Administration)
- Napomena: ovaj projekat, kao i oni prikazani na prethodnom predavanju integriše više izvora podataka.

„Analiza Avionskih Letova na Području Sjedinjenih Američkih Država, Analiza kašnjenja“

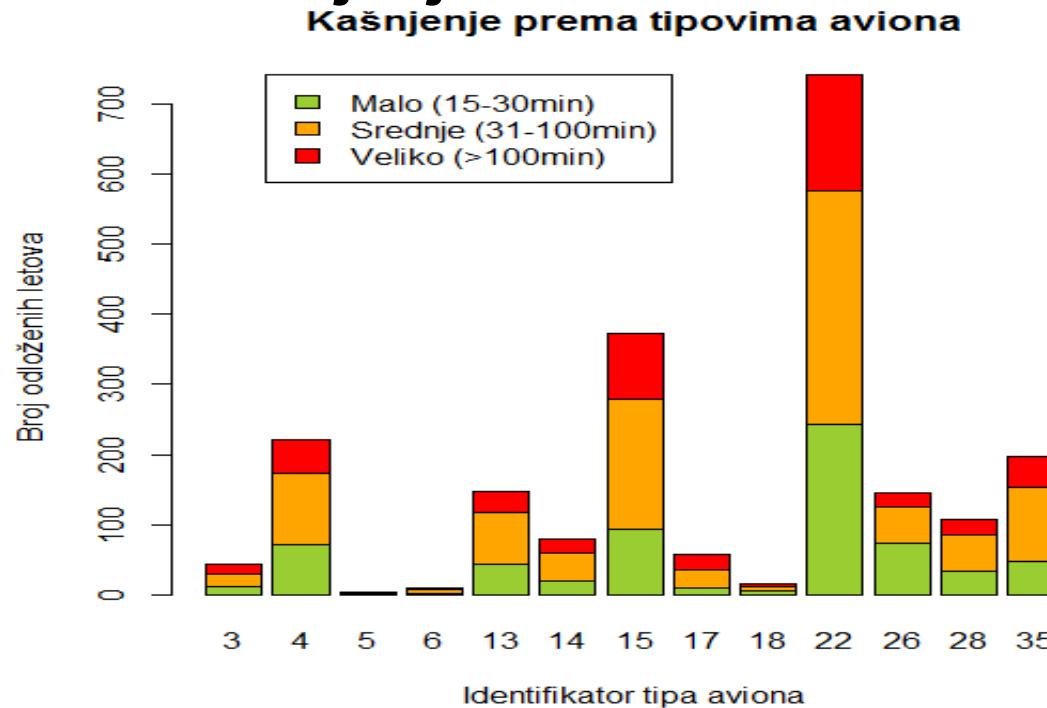
Raspodjela letova po periodima u toku dana



- 1 - Early Morning (03:00AM - 6:59AM)
- 2 - AM Peak (07:00AM - 8:59AM)
- 3 - Late Morning (09:00AM - 11:59AM)
- 4 - Afternoon (12:00PM - 15:59PM)
- 5 - PM Peak (16:00PM - 17:59PM)
- 6 - Evening (18:00PM - 23:59PM)
- 7 - Late Night (00:00AM - 2:59AM)

Letovi su se najčešće odvijali u periodu između 12h i 16h, međutim najviše kašnjenja, zabeleženo je u večernjim satima.

„Analiza Avionskih Letova na Području Sjedinjenih Američkih Država, Analiza kašnjenja“

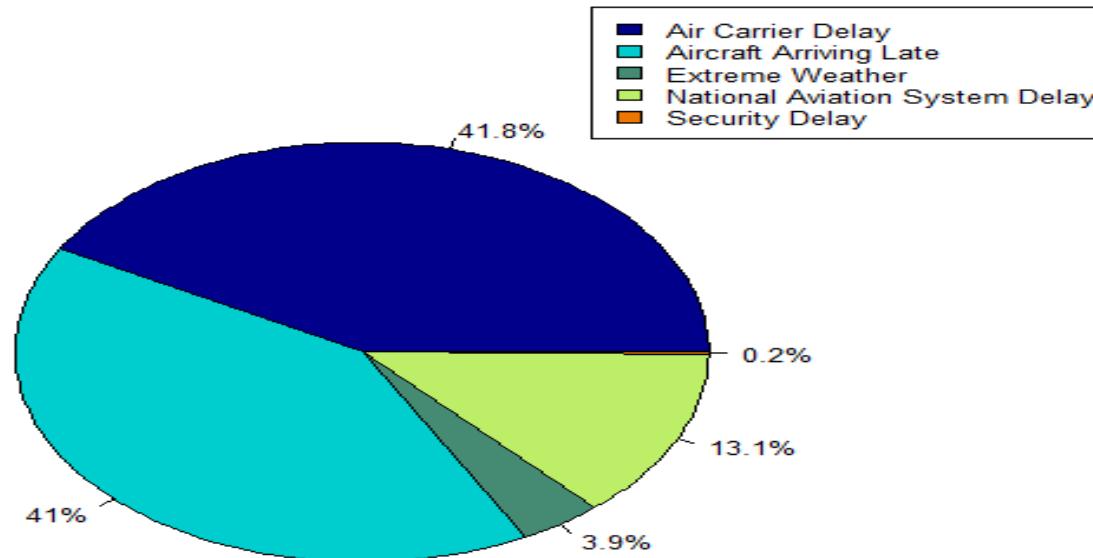


Avion tip 22 – *Canadair Regional Jet* ubedljivo najviše kasni.

Takođe, veliki broj kašnjenja je primećen i kod aviona sa identifikacionim brojem tipa 15 – *Boeing 737 Next Gen.*

„Analiza Avionskih Letova na Području Sjedinjenih Američkih Država, Analiza kašnjenja“

Uzroci odlaganja letova



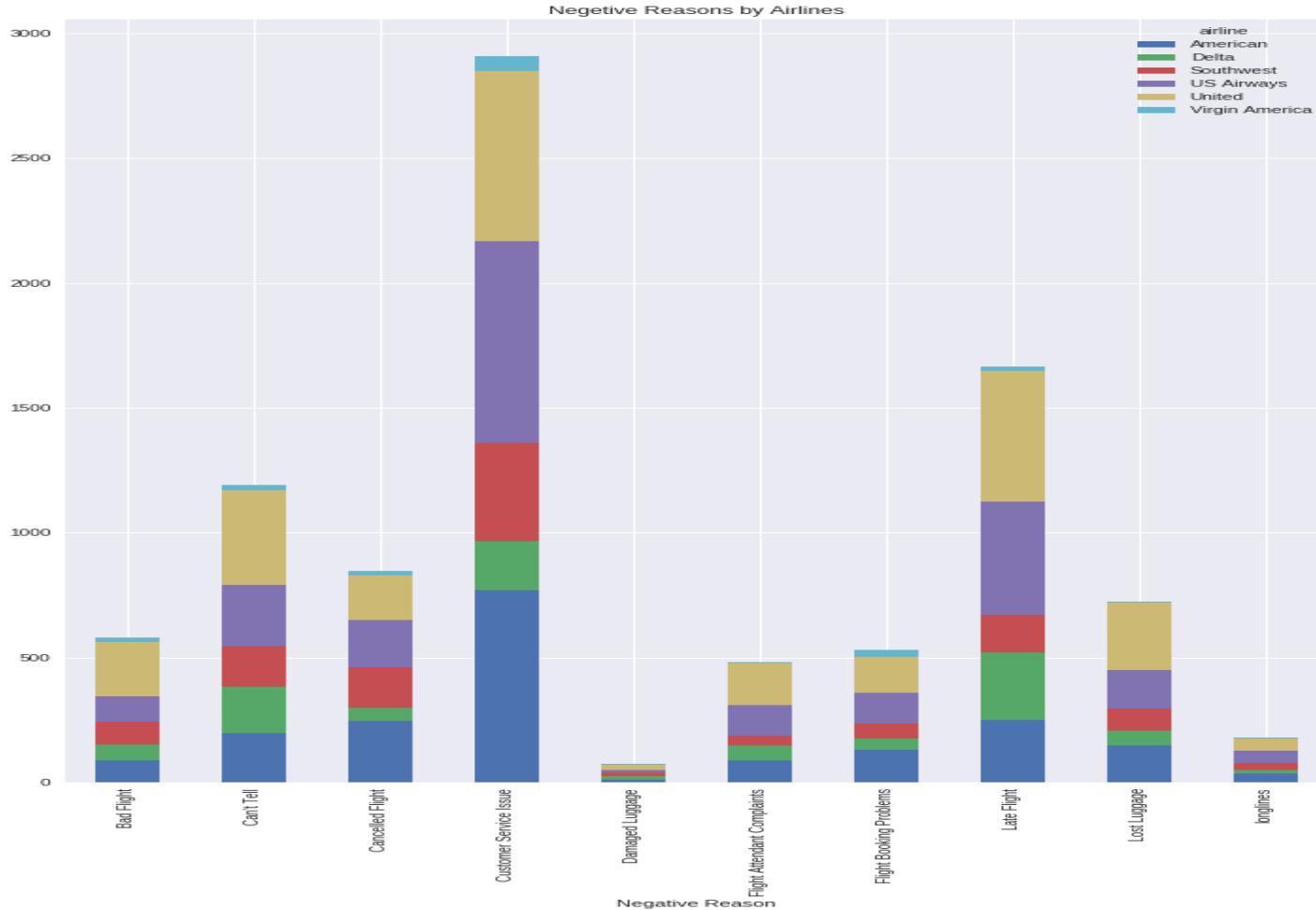
Posebno se izdvajaju kašnjenje za koje je odgovorna avio kompanija (*Air Carrier Delay*) i kašnjenje nastalo kao posledica kasnog dolaska aviona (*Aircraft Arriving Late*)

Interesatno: Samo 3,9 % letova odloženo zbog nepovoljnih vremenskih uslova.

“Sentiment analiza i poređenje metoda: tweet-ovi o letovima američkih avio-kompanija - Utisci korisnika nakon leta”

- Autori: Nikola Đuza, Jana Vojnović, Marina Nenić
- Cilj: analiza sentimenta tweet-ova koje su objavili korisnici nakon putovanja avionom.
- Podaci: 14485 postova objavljenih za mesec Februar, 2015. godine
- Atributi:
 - id, avio kompanija, korisničko ime, datum i vreme, lokacija, vremensku zonu korisnika, tekst tvita, sentiment, ručno utvrđen razlog zbog koga je tvit negativan (ukoliko se može utvrditi)...

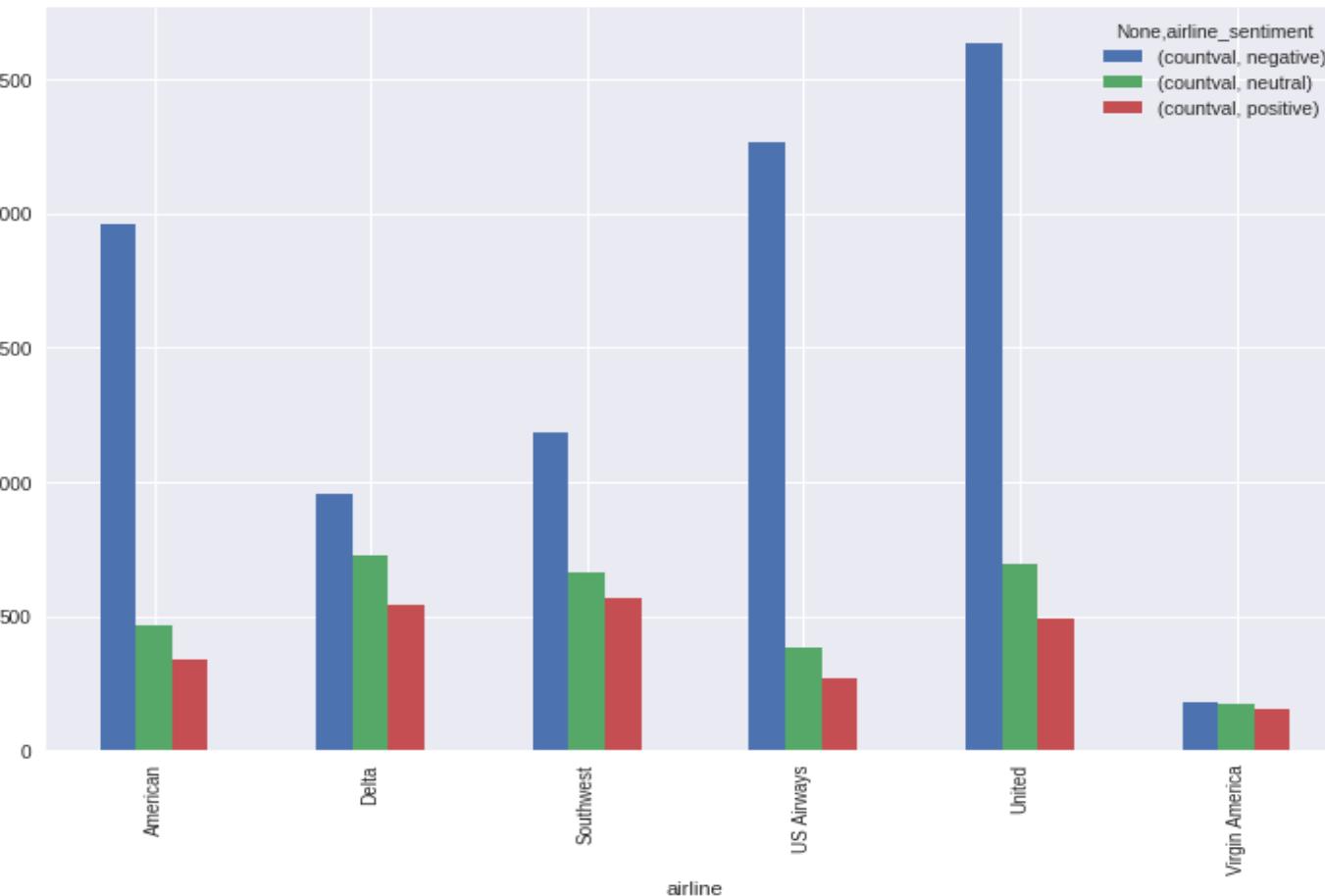
“Sentiment analiza i poređenje metoda: tweet-ovi o letovima američkih avio-kompanija“



Problem sa korisničkim servisom ubedljivo najčešći razlog nezadovoljstva klijenata.

Interesantno:
Iznenadujući je i podatak da je najmanji uzrok nezadovoljstva oštećen prtljag.

“Sentiment analiza i poređenje metoda: tweet-ovi o letovima američkih avio-kompanija“



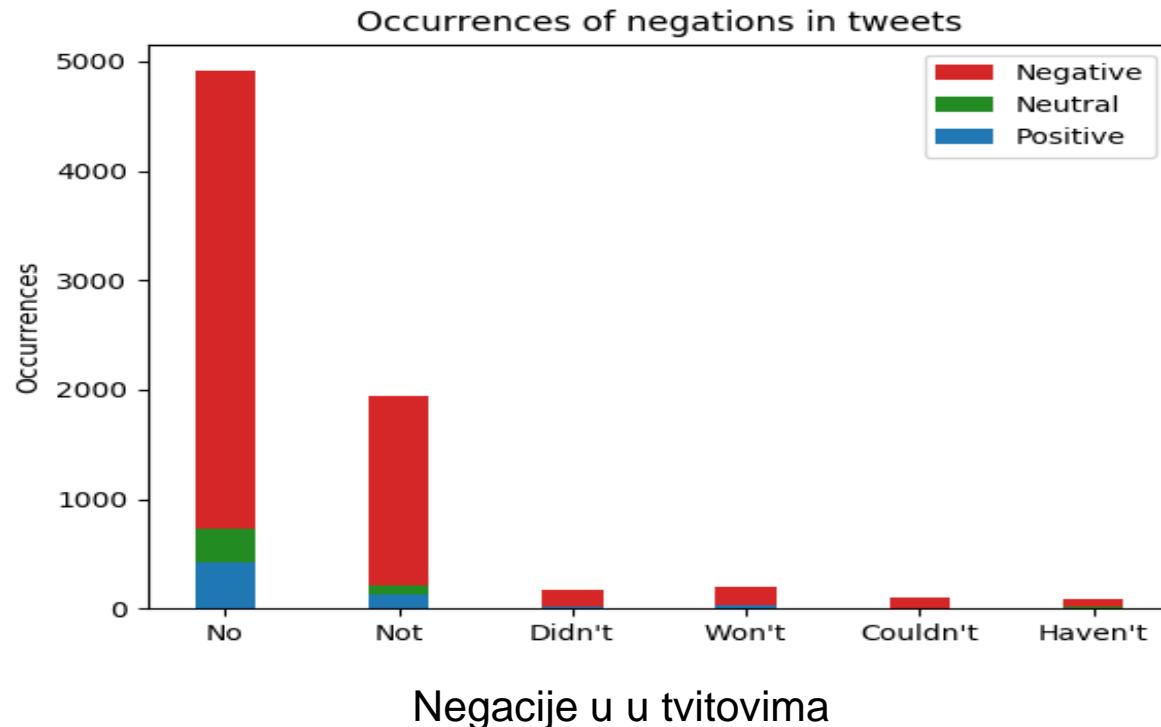
United Airlines zauzima prvo mesto po broju letova, ali i po broju negativnih utisaka.

Nasuprot tome, kompanija Virgin Amerika ima najmanje letova i najmanje negativnih objava. Treba uzeti u obzir i veličine samih kompanija.

Southwest Airlines ima najviše pozitivnih utisaka.

“Sentiment analiza i poređenje metoda: tweet-ovi o letovima američkih avio-kompanija“

❤️ \u2764	31
⌚ \u263a	10
👉 \U0001f44d	42
😊 \U0001f621	33
😢 \U0001f622	32
💕 \U0001f49c	2
✈ \u2708	57
🍷 \U0001f377	2
👉 \U0001f4ba	4
😊 \U0001f60a	31
😢 \U0001f60d	10
👉 \U0001f44c	14
❤️ \U0001f495	6
☀ \U0001f31e	3
😊 \U0001f603	16
😢 \U0001f629	18
😭 \U0001f62d	44
👉 \U0001f60e	6
🎥 \U0001f649	1
😊 \U0001f601	18
❄ \u2744	8
👉 \U0001f44f	28
😭 \U0001f602	62



Emotikoni koji se najčešće javljaju u tvitovima

“Sentiment analiza i poređenje metoda: tweet-ovi o letovima američkih avio-kompanija“



Wordcloud za pozitivne tvitove



Wordcloud za negativne tvitove

„Predviđanje budućih cena akcija“

- Autor: Tijana Sekulić
- Cilj: Uporediti tehnike za analizu vremenskih serija u oblasti finansija
- Podaci: podaci o akcijama Apple od 2010 do 2016 sa <http://finance.yahoo.com/>
- Atributi:
 - Date, Open price, High price, Low price, Close price, Volume, Adjusted Closed price

„Predviđanje budućih cena akcija“

— Adj Close



„Predviđanje budućih cena akcija“

- Neki od zanimljivih uvida u grafik (u terminologiji koju je koristila autorka):
- Od decembra 2010. godine do septembra 2012. godine sa slike vidimo da cena prati rastući trend
- Odnosno za manje od dve godine cene akcije su porasle za oko 50 dolara po akciji.
- U tom periodu kompanija Apple po prvi put od 1989. godine je premašila svog najvećeg konkurenta Microsoft i postala je najveći potrošački brend.
- Jula 2011. godine u kratkom periodu njihove finansijske rezerve su bile veće od Vlade SAD-a.

„Predviđanje budućih cena akcija“

- Predsednik kompanije Apple, Stiv Jobs, preminuo je 5. oktobra 2011. godine.
- Krajem 2012. godine smrt predsednika uzrokovala je privremeni pad cene na oko 53 dolara po akciji.
- Posle jula 2013. godine popularnost i prodaja proizvoda je dostigla vrhunac puštanjem iPhon-a 4 "Steve".
- Cena akcija ulazi u rastući trend sve do juna 2015. godine, a 22. maja 2015. dostiže maksimalnu cenu od 130.671 dolara po akciji.
- Posle jula 2015. godine cene akcije imaju blagi pad.