Analiza socijalnih mreža

Priprema i čišćenje podataka

Marko Mišić, Jelica Protić

13M111ASM

2020/2021.

Uvod (1)

- Podaci za analizu dolaze iz različitih izvora
  - Senzori, upitnici, logovi, rečnici, baze podataka
- Prilikom formiranja skupova za analizu neminovno dolazi do grešaka
  - Greške u unosu od strane korisnika
    - Tipografske greške, nepotpuni unosi, različiti poreci, itd.
  - Korupcija u prenosu ili prilikom smeštanja u skladišta
  - Različite definicije istih entiteta u različitim skladištima
- Potreba da se podaci dovedu u konzistentno i uniformno stanje
  - Kako bi se izbegli pogrešni zaključci pri odlučivanju

Uvod (2)

- Čišćenje podataka (*data cleansing, data cleaning*) je proces otkrivanja, korigovanja ili uklanjanja neispravnih zapisa iz skupa podataka
  - Identifikacija nekompletnih, nekorektnih, netačnih ili nepotrenih delova podataka
  - Ispravljanje ili uklanjanje prljavih i nepotpunih podaka
- Radi se nad primarnim (prljavim) skupom podataka
  - Rezultat je sekundardni skup podataka
    - Skup podataka za analizu
- Sprovodi se ručno ili korišćenjem skripti
  - Obrada u MS Excel ili editorima teksta
  - Procesiranje kroz Python, R ili Perl

Uvod (3)

- Uobičajeno se sastoji od sledećih koraka:
  - Inspekcija podataka i utvrđivanje kvaliteta podataka
  - Čišćenje
  - Verifikacija i validacija
  - Izveštavanje
- Čišćenje podataka se razlikuje od klasične validacije podataka
  - Validacija podataka se obično radi pri unosu
    - Rezultira odbacivanjem unosa, ukoliko podatak nije validan
  - Čišćenje podataka se radi nad većim brojem zapisa

Primer – problem razrešavanja afilijacija (1)

- Primer Elektrotehničkog fakulteta u naučnoj indeksnoj bazi Web Of Science (WoS)
  - Identifikovano 15 različitih načina da se zada afilijacija!
  1. Elektrotehn Fak Beogradu, Belgrade 11120, Serbia
  2. Elektrotehnicki Fak Beogradu, Beograd 11120, Serbia
  3. Fac Elect Engn, Belgrade 11000, Serbia
  4. Fac Elect Engn, Belgrade 11120, Serbia
  5. Sch Elect Engn, Belgrade, Serbia
  6. Univ Belgrade, Dept Comp Engn & Comp Sci, Sch Elect Engn, Belgrade 11001, Serbia
  7. Univ Belgrade, Dept Elect Engn, Belgrade 11000, Serbia

Primer – problem razrešavanja afilijacija (2)

- Nastavak:
  8. Sch Elect Engn ETF, Dept Commun, Belgrade, Serbia
  9. Sch Elect Engn, Belgrade 11120, Serbia
  10. Univ Belgrade, Electrotech Fac, Belgrade 11120, Serbia
  11. Univ Belgrade, Fac Elect Engn, Belgrade 11000, Serbia
  12. Univ Belgrade, Power Converters & Drives Dept, Fac Elect Engn, Beograd 11000, Serbia
  13. Univ Belgrade, Sch Elect Engn, Belgrade 11000, Serbia
  14. Univ Belgrade, Sch Elect Engn, Signals & Syst Dept, Belgrade 11000, Serbia
  15. Univ Beogradu, Elektrotehn Fak, Belgrade 11120, Serbia

Kvalitet podataka

- Podaci moraju da zadovolje određene kriterijume kvaliteta
- Kriterijumi kvaliteta podataka uključuju:
  - Validnost
  - Tačnost
  - Kompletnost
  - Konzistentnost
  - Uniformnost

Validnost podataka (1)

- Zadovoljavanje kriterijuma vezanih za (poslovna) pravila i ograničenja
- Ograničenja tipa podataka
  - Vrednosti određenih kolona moraju biti numeričke, logičke, u obliku datuma i sl.
- Ograničenja opsega
  - Numeričke vrednosti i datumi najčešće moraju biti u određenom opsegu
- Obavezna polja
  - Neke kolone u skupu podataka ne smeju biti prazne
- Ograničenja jedinstvenosti
  - Polja ili kombinacije polja moraju imati jedinstvene vrednosti na celom skupu podataka

Validnost podataka (2)

- Ograničenja vezana za pripadnost skupu vrednosti
  - Vrednosti kolone moraju pripadati ograničenom skupu vrednosti
  - Npr. pol mora biti muški ili ženski
- Ograničenja stranih ključeva
  - Kao strani ključ se ne može koristiti vrednost koja ne postoji kao primarni ključ
- Ograničenja zadata regularnim izrazima
  - Tekstualna polja moraju biti u određenom formatu
  - Npr. polje e-mail adrese ili telefonskog broja
- Unakrsna validacija podataka
  - Određeni uslovi moraju da važe nad više polja
  - Npr. datum izdavanja diplome ne može biti pre datuma upisa

Tačnost podataka

- Stepen usklađenosti podataka sa tačnim vrednostima
- Definisanje mogućih validnih vrednosti ne znači da su te vrednosti tačne
  - Validna poštanska adresa ne mora postojati u realnosti
  - Broj telefona u validnom formatu može biti isključen ili nepostojeći
- Treba napraviti razliku između tačnosti i preciznosti
  - Živeti u Beogradu vs. živeti na Palilulu
- Treba koristiti eksterne baze podataka kako bi se podigla tačnost podataka

Kompletnost podataka

- Stepen potpunosti podataka
  - U kontekstu da su svih neophodni podaci poznati
- Pojavljuje se iz različitih razloga
  - Može se rešiti ispitivanjem originalnog izvora, ukoliko je moguće
- Ponekad se rešava dodavanjem podrazumevanih vrednosti
  - *unknown* ili *missing* vrednosti polja
  - To ne implicira da su podaci potpuni

Konzistentnost podataka

- Podrazumeva usklađenost vrednosti polja ili mera u okviru istog skupa podataka ili u više skupova podataka
- Nekonzistentnost se javlja kada su dve vrednosti u okviru skupa podataka kontradiktorne
  - Na primer, klijent u okviru dva skupa podataka ima dve zabeležene adrese
  - Postavlja se pitanje koji podatak je tačan
    - Da li je tačan onaj koji je kasnije zabeležen?
    - Koji izvor podataka je pouzdaniji?

Uniformnost podataka

- Stepen usklađenosti podataka u kontekstu korišćenja istih jedinica i mera
  - Različiti sistemi mera
  - Načini zapisivanja datuma i vremena
  - Različite novčane jedinice
  - Način izražavanja procenata ili verovatnoća
- Podaci se moraju konvertovati u isti oblik
  - U istu jedinicu mere
- Podrazumeva definisanje normalizovane (kanoničke) forme podataka

Proces čišćenja podataka

- Proces ćišćenja podataka se uobičajeno sastoji od sledećih koraka:
  1. Inspekcija
     - Detekcija neočekivanih, nekorektnih i nekonzistentih podataka
  2. Čišćenje
     - Modifikacija ili uklanjenje problematičnih podataka
  3. Verifikacija
     - Ponavljanje inspekcije da se utvrdi korektnost
  4. Izveštavanje
     - Beleženje načinjenih promena i ocenjivanje kvaliteta
- Uobičajeno iterativni proces dok se ne dobije zadovoljavajući kvalitet podataka

Inspekcija podataka (1)

- Često vremenski zahtevna faza
  - Upotreba softverskih paketa i biblioteka može pomoći
- Profajliranje podataka može pomoći da se utvrdi kvalitet podataka
  - Provera da li kolone zadovoljavaju određene standarde ili šablone
  - Provera tipa podataka
  - Utvrđivanje distribucije pojavljivanja podataka
  - Utvrđivanje jedinstvenih vrednosti u koloni
  - Određivanje broja nepotpunih polja po kolonama

Inspekcija podataka (2)

- Vizuelizacija podataka može pomoći da se identifikuju problematične vrednosti
- Najčešće uz korišćenje statističkih tehnika:
  - Određivanje aritmetičke sredine i standardne devijacije
  - Određivanje opsega i kvantila u kojima se vrednosti nalaze
- Vizuelizacija može pomoći u pronalaženju neobičnih vrednosti (*outliers*)
  - Mogu, a ne moraju da predstavljaju nekorektne podatke

Čišćenje podataka

- Uključuje različite tehnike zavisno od problema i tipa podataka
- Generalno, nekorektni podaci se:
  - Uklanjaju
  - Koriguju
  - Zamenjuju

Nerelevantni podaci

- Relevantni podaci se definišu u kontekstu istraživanja koje se sprovodi
- Na primer:
  - Ukoliko se analiziraju podaci o generalnog prodaji nekog proizvoda, podatak o telefonskom broju klijenta se može zanemariti
    - Na nivou cele kolone
  - Ukoliko se analiziraju podaci o prodaji nekog proizvoda u jednom gradu, podaci o drugim gradovima se mogu zanemariti
    - Na nivou cele vrste
- Podatke treba izostaviti samo ako ne korelišu međusobno

Duplikati podataka

- Često se javljaju:
  - Prilikom kombinovanja podataka iz više izvora
  - Greškom korisnika:
    - Kada dva puta pritisne dugme za predaju forme
    - Drugi put popuni isti obrazac da bi korigovao podatke
- Bitno je jednoznačno ih identifikovati nekim poljem sa osobinama primarnog ključa
- Obično se uklanjaju

Konverzije tipova

- Numeričke vrednosti bi trebalo konvertovati u odgovarajuće numeričke tipove
  - Ne čuvati u obliku stringa
- Kategoričke vrednosti se mogu konvertovati u brojeve ili obratno
  - Na primer, FALSE $\rightarrow$ 0, TRUE $\rightarrow$ 1 i obratno
- Konverzije mogu proizvesti greške, ukoliko vrednost ne može da se konvertuje u zadati tip
  - Treba da se dobiju NA (*Not Available*) vrednosti
  - Ispravka i ispitivanje zbog čega

Sintaksne i slične greške (1)

- Najčešće proizvodi korisnik prilikom unosa
- Uklanjanje suvišnih belih znaka
  - Podrazumeva i zamenu tabulacija jednim belim znakom
- Dopunjavanje stringova do određene širine
- Veći broj načina za unos istog stringa
- Npr. unos podataka o polu
  - Gender: m, Male, fem., FemalE, Femle
    - Kategorička promenljiva sa dve vrednosti, ali pet identifikovanih klasa

Sintaksne i slične greške (2)

- Tri načina za razrešavanje grešaka
- Korišćenje rečnika
  - Podrazumeva mapiranje svih vrednosti u korektne vrednosti nakon identifikacije svih mogućih klasa
- Korišćenje regularnih izraza
  - U prethodnom primeru, sve reči koji počinju sa *m* bi se mapirale u *male*, a sve koje počinju sa *f* u *female*
- Aproksimativno mapiranje
  - Zasnovano na nekoj meri distance od korektne forme
    - Levenštajnova (*edit*) distanca

Standardizacija podataka

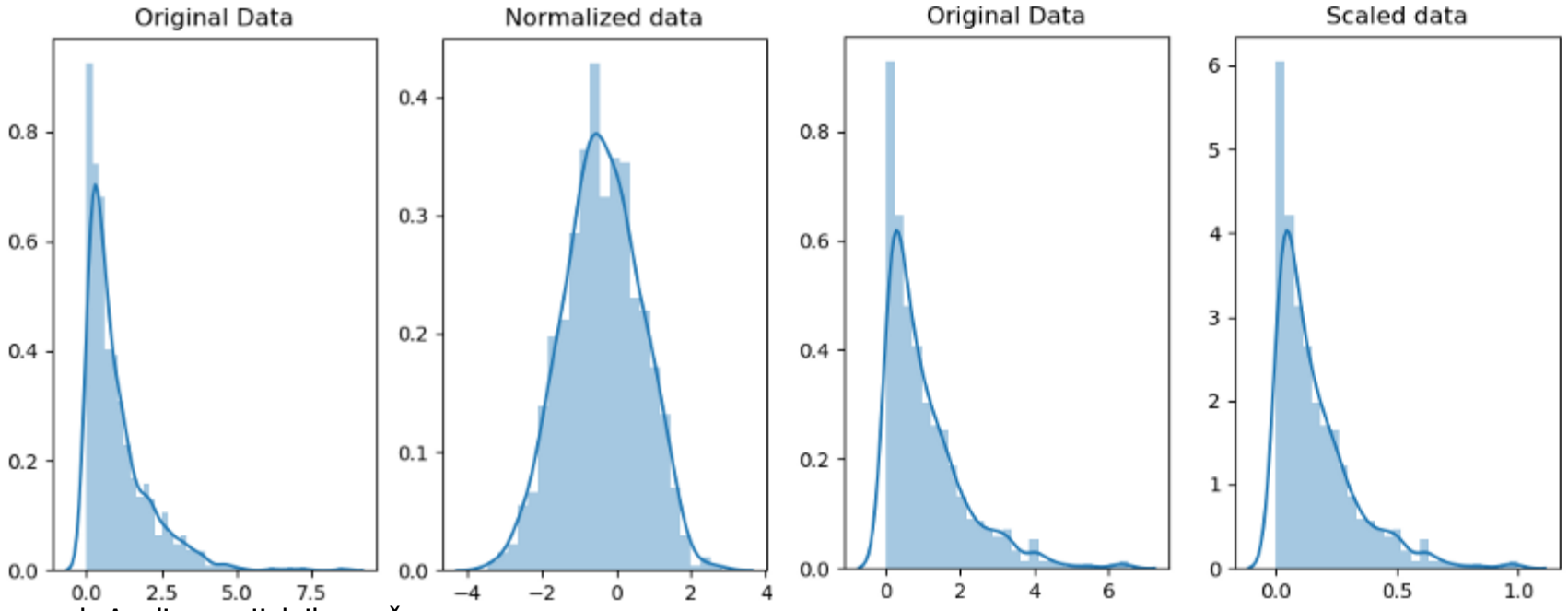
- Pretvaranje svih vrednosti u isti, standardizovani format zavisno od tipa podataka
  - Svođenje na istu, kanoničku formu
- Stringovi se pretvaraju u sva mala ili sva velika slova
- Numeričke vrednosti se pretvaraju u iste jedinice mere
- Datumi se pretvaraju u isti format
  - USA ili EU
  - Korišćenje *timestamp-a* (u milisekundama) nije isto kao korišćenje datumskih objekata

Skaliranje i normalizacija podataka (1)

- Svođenje (numeričkih) podataka na odgovarajuću (sličnu) skalu
  - Najčešće su u pitanju podaci u pokretnom zarezu
  - Na primer: 0-1, 0-100, -1 do 1
- Često se radi i transformacija podataka tako da budu u skladu sa normalnom (Gausovom) raspodelom
  - Za one metode analize koji očekuju normalnu raspodelu podataka
- Omogućava poređenje različitih vrednosti
- Olakšava i vizuelizaciju podataka
  - Uz korišćenje inverzne ili logaritamske skale i sl.

Skaliranje i normalizacija podataka (2)

- Različite tehnike skaliranja i normalizacije podataka
- Na primer, skaliranje prosečne ocene studenta:
  - USA GPA 0-5
  - SRB prosečna ocena 6-10

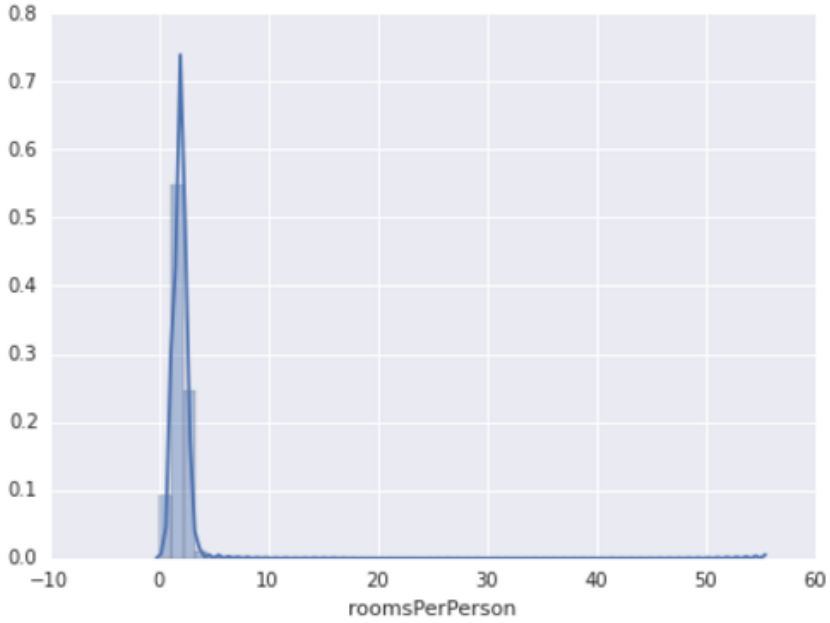


Skaliranje i normalizacija podataka (3)

- Skaliranje u opsegu (*scaling to a range*)
  - Konvertovanje podataka iz prirodnog opsega u zadati opseg
  - Pogodno kada su poznate donja i gornja granica opsega
    - Nema mnogo ekstremnih vrednosti
    - Uniformna je distribucija podataka
- Dobar primer: podaci o godištu ljudi
  - Velika većina ljudi je starosti 0-90 godina
  - Relativno uniformna raspodela
- Loš primer: podaci o prihodu ljudi
  - Veliki broj ljudi ima male prihode
  - Relativno mali broj ljudi ima jako velike prihode

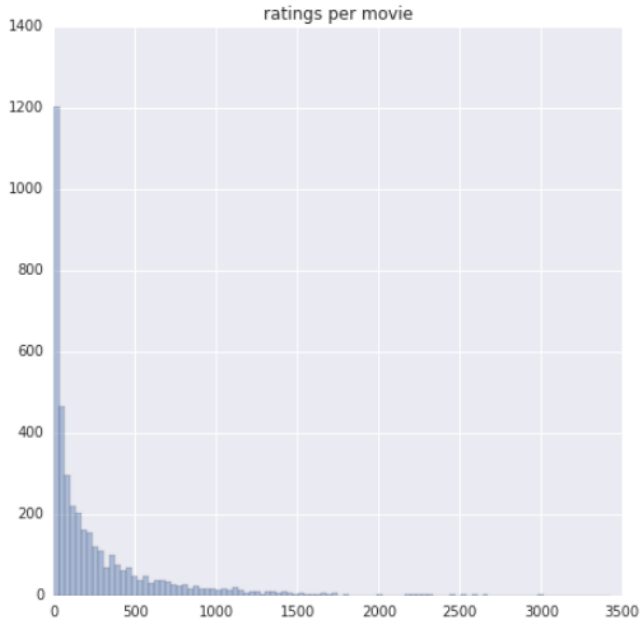
Skaliranje i normalizacija podataka (4)

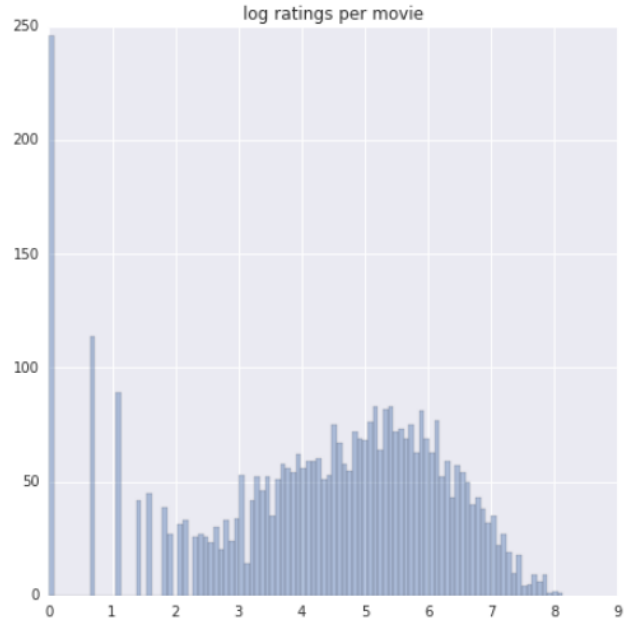
- Odsecanje odlika (*feature clipping*)
  - Sve ekstremne vrednosti se postavljaju na neku unapred definisanu maksimalnu vrednost
  - Formiraju se „brežuljci“ na krajevima opsega



Skaliranje i normalizacija podataka (5)

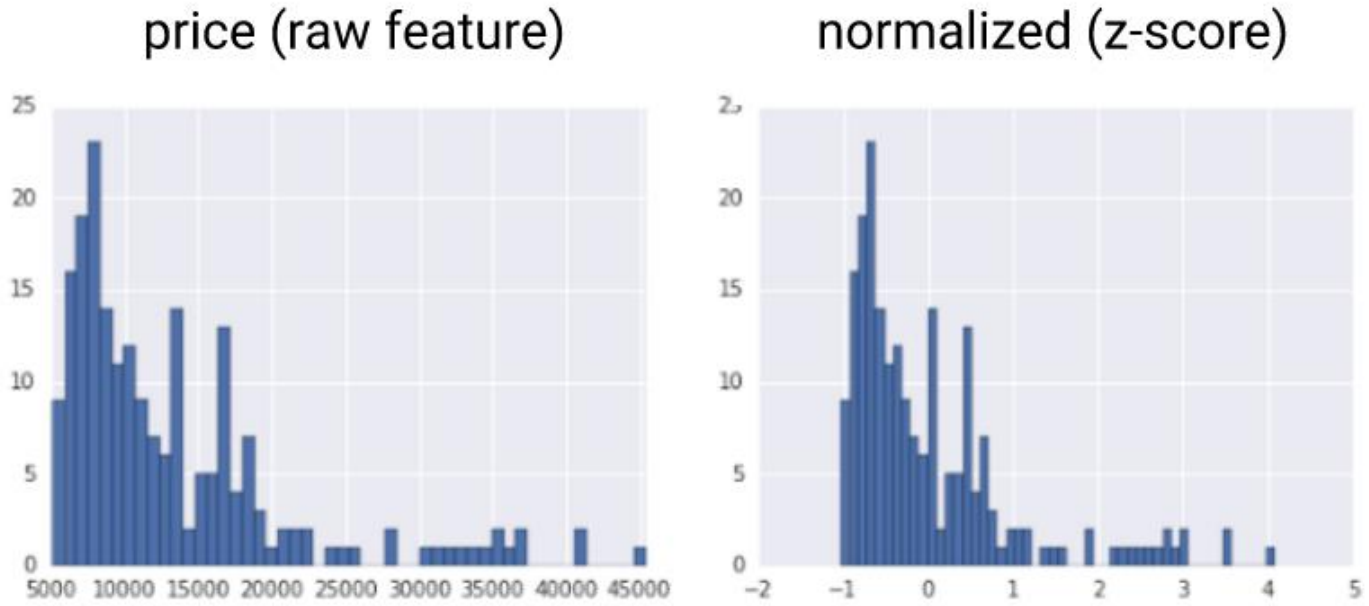
- Logaritamsko skaliranje (*log scaling*)
  - Vrednost se zamenjuje sopstvenim logaritmom
  - Pogodno kod podataka koji prate *power law* distribuciju
  - Primer – rejtinzi filmova





Skaliranje i normalizacija podataka (6)

- Z-skor ili standardni skor (*z-score*)
  - Bazirano na aritmetičkoj sredini i standardnoj devijaciji
  - Predstavlja broj standardnih devijacija za koje je sirov skor udaljen od srednje vrednosti



Nedostajuće vrednosti (1)

- Tri najčešća načina za rešavanje problema nedostajućih (*missing*) vrednosti
- Izostavljanje vrste ili kolone
  - Ukoliko se nedostajuća vrednost javlja retko
  - Najčešće kod statističkih analiza
    - Druge tehnike mogu pokvariti rezultate
- Zamena (izračunavanje) nedostajuće vrednosti
  - Izračunavanje se vrši na osnovu preostalih vrednosti

Nedostajuće vrednosti (2)

- Statistička zamena:
  - Srednja vrednost
  - Medijana (robusnija)
- Zamena korišćenjem linearne regresije
  - Osetljivo na pojavu *outlier*-a
- Zamena na osnovu sličnih zapisa
  - Samo kada postoji dovoljno podataka
  - Za numeričke i kategoričke podatke
  - Na slučajan način (*random*)
  - Na osnovu sortiranja po pomoćnim promenljivama
  - *k-nearest neighbours* zamena

Nedostajuće vrednosti (3)

- Obeležavanje nedostajućih vrednosti
  - Kada se nedostajuće vrednosti ne javljaju na slučajan način
  - Popunjavanje nedostajućih vrednosti nekom karakterističnom vrednošću
    - Npr. vrednost 0
  - Ne uzima se u obzir prilikom računanja statističkih pokazatelja
- Nepostojanje vrednosti samo po sebi nosi određenu informaciju

Ekstremne vrednosti (*outliers*)

- Podaci koji značajno odstupaju od ostalih podataka u skupu
- Mogu biti rezultat lošeg prikupljanja podataka, ali i korektni podaci koji predstavljaju važne anomalije u skupu podataka
- Ne treba ih uklanjati, osim ako postoji dobar razlog
  - Prethodno dobro istražiti da li predstavljaju neku naročitu osobinu skupa podataka

Greške u zapisima

- Mogu postojati na nivou jednog zapisa ili u okviru različitih zapisa jednog skupa podataka
  - Unose kontradiktornost, odnosno nekonzistentnost
  - Npr. ukoliko neka suma ne odgovara zbiru pojedinačnih kolona
- Mogu se razrešiti, ukoliko postoje adekvatni podaci u preostalim kolonama

Verifikacija podataka

- Nakon završetka čišćenja,
  podatke je potrebno proveriti još jednom
  - Provera pravila i ograničenja
- Na primer,
  nakon popunjavanja nedostajućih podataka,
  može se desiti da popunjene vrednosti
  narušavaju postavljena ograničenja
- Često uključuje ručne ispravke

Izveštavanje i ocenjivanje kvaliteta

- Beleženje svih akcija koje su izvršene nad podacima
  - Koje izmene su načinjene
  - Koja pravila su i koliko puta narušena
- Često je podržano od strane korišćenih softverskih paketa
  - Nije teško ugraditi ni logiku za logovanje u okviru skripti za čišćenje

Literatura

- O. El Gabry, The Ultimate Guide to Data Cleaning, https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
- E. Rahm, H. H. Do, Data Cleaning: Problems and Current Approaches, IEEE Data Eng. Bull. 1999.
- Data Preparation and Feature Engineering for Machine Learning, https://developers.google.com/machine-learning/data-prep/transform/normalization
- M. Popovic, I. Mitrovic, J. Protic, Problemi sa afilijacijama i njihov uticaj na rangiranje univerziteta, XX Trend, Kopaonik, 2014.
- Data cleansing, https://en.wikipedia.org/wiki/Data_cleansing