Istraživanje podataka

Vežbe 1

19. Februar 2021

Outline

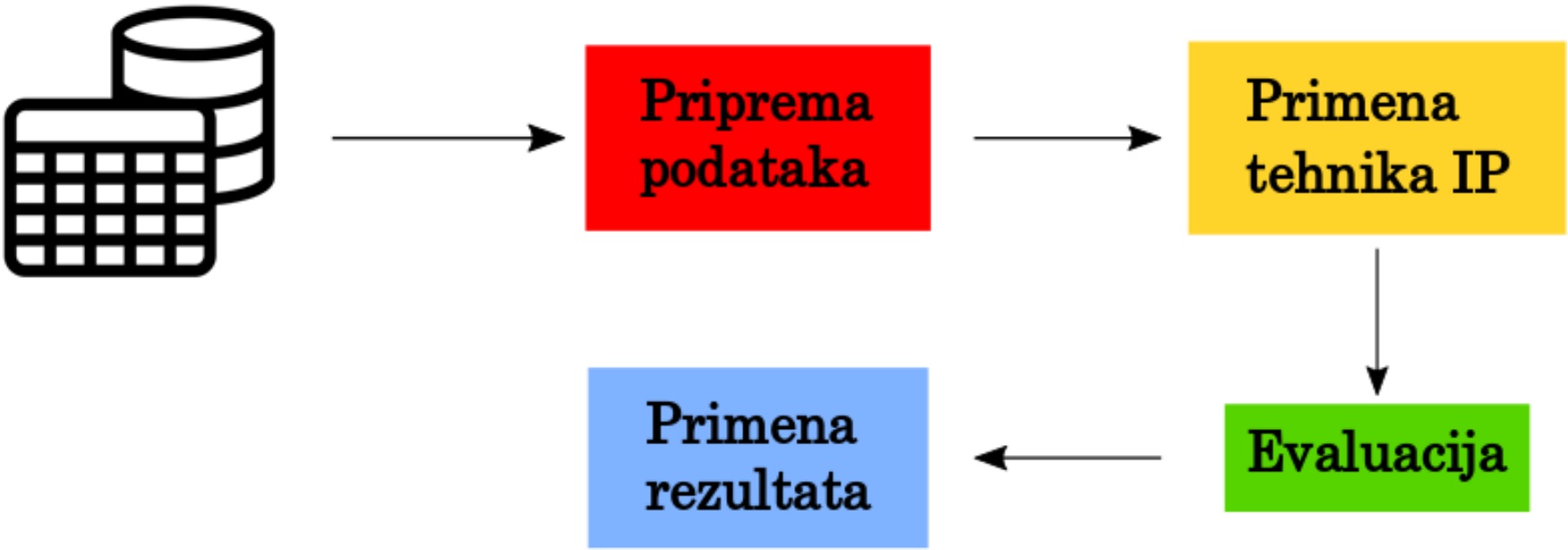
1. Istraživanje podataka
2. Skupovi podataka i atributi
3. Šum i elementi van granice
4. IBM SPSS Modeler
5. Učitavanje podataka
6. Zadatak

Outline

1. Istraživanje podataka
2. Skupovi podataka i atributi
3. Šum i elementi van granice
4. IBM SPSS Modeler
5. Učitavanje podataka
6. Zadatak

Istraživanje podataka

Istraživanje podataka je proces automatskog otkrivanja korisnih informacija u velikom skladištu podataka.



Razumevanje istraživanja podataka

Pitanja na koja je potrebno odgovoriti pri planiranju istraživanja:

- Koji problem želite da rešite?
- Koji izvori podataka su dostupni i koji su delovi podataka bitni za trenutni problem?
- Koju vrstu pretprocesiranja morate da uradite pre nego što počnete da koristite podatke?
- Koju tehniku/tehnike istraživanja podataka ćete koristiti?
- Kako ćete proceniti rezultate analize podataka?
- Kako ćete dobiti najveću korist od informacija koje ste dobili istraživanjem podataka?

CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) - metodologija koja se pokazala uspešnom u industriji.

Faze

1. ***Razumevanje posla*** Utvrđivanje poslovnih ciljeva, određivanje ciljeva istraživanja podataka.
2. ***Razumevanje podataka*** Podaci obezbeđuju *sirovine* za istraživanje podataka. Ova faza se bavi upoznavanjem izvora podataka i njihovih karakteristika. Uključuje prikupljanje početnih podataka, opisivanje podataka, upoznavanje podataka i proveru kvaliteta podataka.

CRISP-DM

Faze

3 ***Priprema podataka*** Odabir, čišćenje, konstruisanje, formatiranje podataka.

4 ***Modeliranje*** Korišćenje metoda za analizu i dobijanje informacija iz podataka. Ova faza uključuje odabir tehnika, izgradnju i procenu modela.

5 ***Evaluacija*** Procena rezultata i određivanje narednih koraka.

6 ***Razvoj*** Integracija novih znanja u svakodnevne poslovne procese kako bi se rešio originalni poslovni problem.

Outline

1 Istraživanje podataka

2 Skupovi podataka i atributi

3 Šum i elementi van granice

4 IBM SPSS Modeler

5 Učitavanje podataka

6 Zadatak

atributi

- Skup podataka - kolekcija objekata (slogova, uzoraka, entiteta...)
- Atributi - svojstvo ili karakteristike objekata
- Vrednosti atributa - brojevi ili simboli koji su pridruženi atributu

Podaci u relacionim bazama podataka

	INDEKS	IME	PREZIME	MESTORODJENJA	DATUPISA
1	20150220	Урош	Рачић	Београд (Савски венац)	06.07.2015
2	20150063	Радмила	Букуров	Београд (Савски венац)	06.07.2015
3	20150352	Анђела	Лазић	Београд (Земун)	06.07.2015
4	20150277	Марија	Марков	Београд (Звездара)	06.07.2015
5	20150037	Матија	Млађеновић	Крушевац	06.07.2015

```
create table Dosije (
indeks integer,
ime varchar(100) ,
prezime varchar(100),
mestorodjenja
varchar(100),
datupisa date
)
```

Podela atributa prema osobinama i operacijama koje mogu da se primene

Za podelu se koriste operacije:

1. Različitost: $=$ i $\neq$
2. Uređenje: $<$, $\leq$, $>$ i $\geq$
3. Aditivnost: $+$ i $-$
4. Multiplikativnost: $*$ i $/$

Podela atributa prema osobinama i operacijama koje mogu da se primene

- Kvalitativni
  - Imenski (eng. Nominal) operacije: 1
  - Redni (eng. Ordinal) operacije: 1,2
- Kvantitativni
  - Intervalni (eng. Interval) operacije: 1,2,3
  - Razmerni (eng. Ratio) operacije: 1,2,3,4

Podela atributa prema osobinama i operacijama koje mogu da se primene

Tip atributa	Opis	Primeri
Imenski (eng. Nominal)	Vrednost imenskog atributa su upravo različita imena, tj. imenski atributi pružaju samo mogućnost razlikovanja jednog od drugog objekta (=, ≠)	poštanski kodovi, identifikacije zaposlenih, boja očiju, pol (muški, ženski)
Redni (eng. Ordinal)	Vrednosti rednih atributa pružaju dovoljno informacija za uređenje objekata (<, >)	redni brojevi zgrada u ulici
Intervalni (eng. Interval)	Za intervalne atribute, ima smisla razlika između vrednosti, tj. postoji jedinica mere takvih atributa (+, -)	datumi u kalendaru
Razmerni (eng. Ratio)	Kod razmernih atributa ima smisla i proizvod i količnik (*, /) tih atributa	količina novca, godine, masa, dužina

Podela atributa prema broju vrednosti koji sadrže

- Diskretni atributi
  - Imaju konačan ili prebrojivo beskonačan skup vrednosti
  - Binarni atributi su specijalan slučaj diskretnih atriubuta
- Kontinuirani (neprekidni) atribututi
  - Skup vrednosti ovih atributa čine realni brojevi

Asimetrični (retki) podaci

- Jedino se prisustvo ne-nula vrednosti smatra značajnim
- Binarni atributi kod kojih su bitne ne-nula vrednosti se zovu asimetrični binarni atributi

Zadaci

1 Za sledeće atribute odrediti da li su binarni, diskretni ili neprekidni. Takođe odrediti da li su kvalitativni (imenski ili redni) ili kvantitativni (intervalni ili razmerni).

- starost u godinama

Zadaci

1 Za sledeće atribute odrediti da li su binarni, diskretni ili neprekidni. Takođe odrediti da li su kvalitativni (imenski ili redni) ili kvantitativni (intervalni ili razmerni).

- starost u godinama
  diskretan, kvantitativni, razmerni

Zadaci

1 Za sledeće atribute odrediti da li su binarni, diskretni ili neprekidni. Takođe odrediti da li su kvalitativni (imenski ili redni) ili kvantitativni (intervalni ili razmerni).

- starost u godinama
  diskretan, kvantitativni, razmerni
- Vreme u oznakama AM ili PM

Zadaci

1 Za sledeće atribute odrediti da li su binarni, diskretni ili neprekidni. Takođe odrediti da li su kvalitativni (imenski ili redni) ili kvantitativni (intervalni ili razmerni).

- starost u godinama
  diskretan, kvantitativni, razmerni
- Vreme u oznakama AM ili PM
  binaran, kvalitativni, redni

Zadaci

- osvetljenost merena ljudskom procenom

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima
  neprekidan, kvantitativan, razmerni

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima
  neprekidan, kvantitativan, razmerni
- bronzane, srebrne i zlatne medalje osvojene na Olimpijadi

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima
  neprekidan, kvantitativan, razmerni
- bronzane, srebrne i zlatne medalje osvojene na Olimpijadi
  diskretan, kvalitativan, redni

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima
  neprekidan, kvantitativan, razmerni
- bronzane, srebrne i zlatne medalje osvojene na Olimpijadi
  diskretan, kvalitativan, redni
- broj pacijenata u bolnici

Zadaci

- osvetljenost merena ljudskom procenom
  diskretan, kvalitativan, redni
- uglovi mereni u stepenima
  neprekidan, kvantitativan, razmerni
- bronzane, srebrne i zlatne medalje osvojene na Olimpijadi
  diskretan, kvalitativan, redni
- broj pacijenata u bolnici
  diskretan, kvantitativan, razmeran

Zadaci

- ISBN brojevi knjiga

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima:
  neproziran, delimično providan (prozračan), transparentan

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci
  diskretan, kvalitativan, redni

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci
  diskretan, kvalitativan, redni
- rastojanje od centra kampusa

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci
  diskretan, kvalitativan, redni
- rastojanje od centra kampusa
  neprekidan, kvantitativan, razmerni

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci
  diskretan, kvalitativan, redni
- rastojanje od centra kampusa
  neprekidan, kvantitativan, razmerni
- broj u garderobi

Zadaci

- ISBN brojevi knjiga
  diskretan, kvalitativan, imenski
- sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan
  diskretan, kvalitativan, redni
- rang u vojsci
  diskretan, kvalitativan, redni
- rastojanje od centra kampusa
  neprekidan, kvantitativan, razmerni
- broj u garderobi
  diskretan, kvalitativan, imenski

Tipovi skupova podataka

- Slogovi
  - Matrica podataka
    - skup numeričkih atributa
  - Podaci u dokumentima
    - atributi istog tipa, asimetrični
  - Transakcioni podaci
    - transakcija (objekat) - skup stavki
- Grafovi
- Podaci sa poretkom (eng. Ordered)
  - Prostorni podaci
  - Vremenski (zavisni) podaci
  - Redosledni podaci

Zadaci

2 Koja veličina ima veću prostornu autokorelaciju: dnevna količina padavina ili dnevna temperatura?

Zadaci

2 Koja veličina ima veću prostornu autokorelaciju: dnevna količina padavina ili dnevna temperatura?
dnevna temperatura

Zadaci

3 Zašto je matrica terma u dokumentima primer skupa podataka koji ima asimetrične diskretne ili asimetrične neprekidne osobine (atribute)?

Zadaci

3 Zašto je matrica terma u dokumentima primer skupa podataka koji ima asimetrične diskretne ili asimetrične neprekidne osobine (atribute)?

U *i.* redu i *j.* koloni matrice čuva se broj poljavljivanja *j.* terma u *i.* dokumentu. Kako većina dokumenata sadrži mali deo svih mogućih reči, 0 vrednosti, koje nemaju značaja u opisu i poređenju dokumenata, će se pojavljivati u velikom broju. Zato matrica ima asimetrične diskretne osobine. Ako se upotrebi normalizacija nad termima i dokumentima, onda matrica ima asimetrične neprekidne atribute.

Outline

1 Istraživanje podataka

2 Skupovi podataka i atributi

3 Šum i elementi van granice

4 IBM SPSS Modeler

5 Učitavanje podataka

6 Zadatak

Šum i elementi van granice

- Šum predstavlja modifikaciju originalnih vrednosti
- Elementi van granica su objekti sa karakteristikama koje su značajno različite od najvećeg broja objekata u skupu podataka

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?
  Ne

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?
  Ne
- Da li su elementi van granica uvek objekti koji spadaju u šum?

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?
  Ne
- Da li su elementi van granica uvek objekti koji spadaju u šum?
  Ne

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?
  Ne
- Da li su elementi van granica uvek objekti koji spadaju u šum?
  Ne
- Da li šum može da pretvori očekivanu vrednost u neobičnu i obrnuto?

Zadaci

4 Napraviti razliku između šuma i elemenata van granica.

- Da li je šum interesantan ili poželjan? Elementi van granica?
  Šum - nije, elementi van granica - jesu
- Da li objekti koji spadaju u šum mogu biti elementi van granica?
  Da
- Da li su objekti koji spadaju u šum uvek elementi van granica?
  Ne
- Da li su elementi van granica uvek objekti koji spadaju u šum?
  Ne
- Da li šum može da pretvori očekivanu vrednost u neobičnu i obrnuto?
  Da

Outline

1 Istraživanje podataka

2 Skupovi podataka i atributi

3 Šum i elementi van granice

4 IBM SPSS Modeler

5 Učitavanje podataka

6 Zadatak

Rad sa podacima u IBM SPSS Modeleru

- Učitavanje podataka
- Manipulacija podacima
- Izvoz rezultata

Rad sa podacima u IBM SPSS Modeleru

- SPSS - Statistical Package for the Social Sciences
- Operacije koje se mogu primeniti nad podacima su predstavljene kao čvorovi. Niz povezanih operacija (čvorova) se naziva tok podataka (eng. data stream). Vezama između čvorova određuje se pravac toka podataka.
- U okviru palete čvorova mogu se izabrati željene operacije (čvorovi).
- Jezičak CRISP-DM obezbeđuje organizaciju projekta prema metodologiji koja se pokazala uspešnom u industriji.

Outline

1 Istraživanje podataka

2 Skupovi podataka i atributi

3 Šum i elementi van granice

4 IBM SPSS Modeler

5 Učitavanje podataka

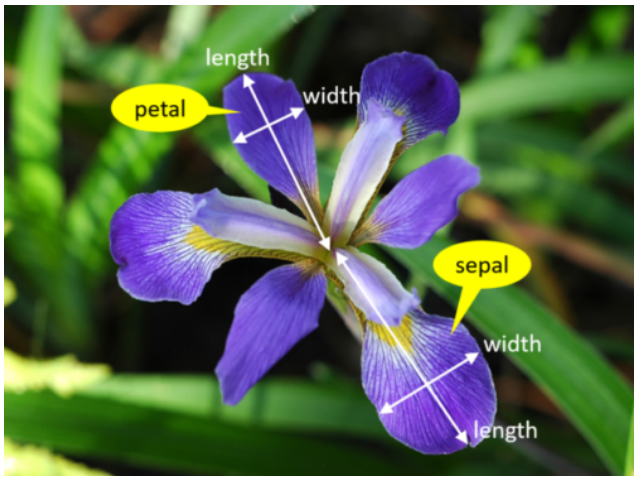
6 Zadatak

Baza podataka za potrebe kursa - ip2019

- konekcija: student/abcdef
- tabele u šemi student:
  - iris - podaci o perunikama;
  - adult - podaci dobijeni pri popisu; koriste se za predviđanje zarade;
  - market_basket - podaci o potrošačkim korpama
- tabele u šemi pekara - podaci o radu pekare

Atributi skupa podataka *iris*

- *sepalwidth* - širina čašičnih listića
- *sepallength* - dužina čašičnih listića
- *petalwidth* - širina latica
- *petalength* - dužina latica
- *class* - klasa



Učitavanje podataka iz baze podataka

- Čvor *Database*
- koristi ODBC (Open Database Connectivity)

Učitavanje podataka iz baze podataka

- opcije:
  - *Data* - učitavanje podataka iz tabele ili rezultata upita
  - *Filter* - odabir atributa
  - *Types* - informacije o atributima
    - *Measurement levels* - tip
    - *Values* - interval ili moguće vrednosti atributa
    - *Missing* - definisanje načina obrade *nedostajućih* vrednosti
    - *Check* - definisanje akcije za objekte koji imaju vrednost koja ne pripada definisanom intervalu ili listi mogućih vrednosti u *Values*

Učitavanje podataka iz baze podataka

- *Measurement levels* - "tip upotrebe "
  - Default - nepoznat, najčešće jer još nije pročitan
  - Continuous - neprekidan, numerički
  - Categorical - kategorički; nakon čitanja mogućih vrednosti prelazi u Flag, Nominal, ili Typeless
  - Flag - binarni
  - Nominal - imenski
  - Ordinal - redni
  - Typeless - za atribute koji imaju jednu vrednost, imenske atribute sa više vrednosti od dozvoljenog broja (Dozvoljeni broj se može promeniti sa koracima: File -> Stream Properties -> Options -> Maximum members for nominal fields)

Učitavanje podataka iz baze podataka

- *Values* - interval ili moguće vrednosti atributa
  - $< Read >$ - informacije se učitavaju pri izvršavanju čvora
  - $< Read+ >$ - informacije se učitavaju i dodaju definisanim (ako postoje)
  - $< Pass >$ - Ne učitavaju se informacije
  - $< Current >$ - ostaju definisane vrednosti
  - Specify... - otvara se poseban prozor za definisanje vrednosti

Učitavanje podataka iz baze podataka

- *Check*
  - None - ne menja se vrednost (podrazumevana akcija)
  - Nullify - postavlja se na null
  - Coerce - vrednost će biti prebačena u legalnu
    - za Flag - u netačnu vrednost
    - Nominal i Ordinal - u prvu vrednost iz skupa
    - za neprekidne - ako je vrednost veća od gornje granice biće zamenjena sa gornjom granicom, a ako je vrednost manja od donje granice biće zamenjena sa donjom granicom zadatog intervala mogućih vrednosti
    - null vrednost za neprekidne atribute se zamenjuje sa srednjom vrednošću

Učitavanje podataka iz baze podataka

- *Check*
  - Discard - ceo slog se odbacuje
  - Warn - broj slogova se napravilnim vrednostima se prijavluje
  - Abort - kada se naiđe na prvi nepravilan slog prijavljuje se greška

Upoznavanje sa podacima

- Data Audit - čvor za upoznavanje sa podacima
  - prikazuje sumarne statistike za atribute i grafike sa distribucijom vrednosti po atributima
  - prikazuje izveštaj o nedostajućim vrednostima, elementima van granica, ekstremnim vrednostima i omogućava definisanje akcija za obradu tih vrednosti

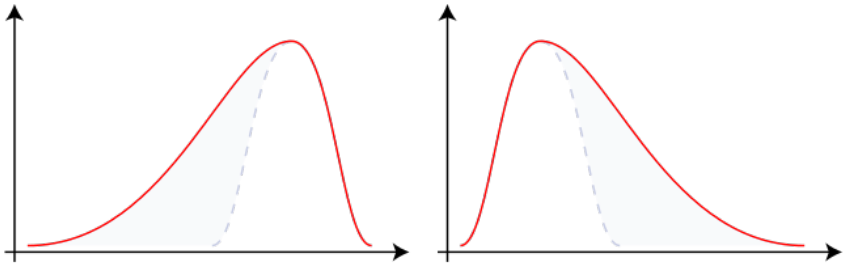
Statistike

Za uzorak od n vrednosti, $x_1, x_2, ..., x_n$

- srednja vrednost (eng. mean) je $\mu = \frac{1}{n} \sum_i x_i$
- varijansa (eng. variance) je $\sigma^2 = \frac{1}{n-1} \sum_i (x_i - \mu)^2$
- standardna devijacija (eng. standard deviation), mera disperzije oko srednje vresnosti $\sigma = \sqrt{\frac{1}{n-1} \sum_i (x_i - \mu)^2}$

Statistike

- iskrivljenost (eng. skewness) Mera asimetrije distribucije. Normalna distribucija je simetrična i ima vrednost asimetrije 0. Distribucija sa značajnom pozitivnom asimetrijom ima dugi desni rep. Distribucija sa značajnom negativnom asimetrijom ima dugačak levi rep.



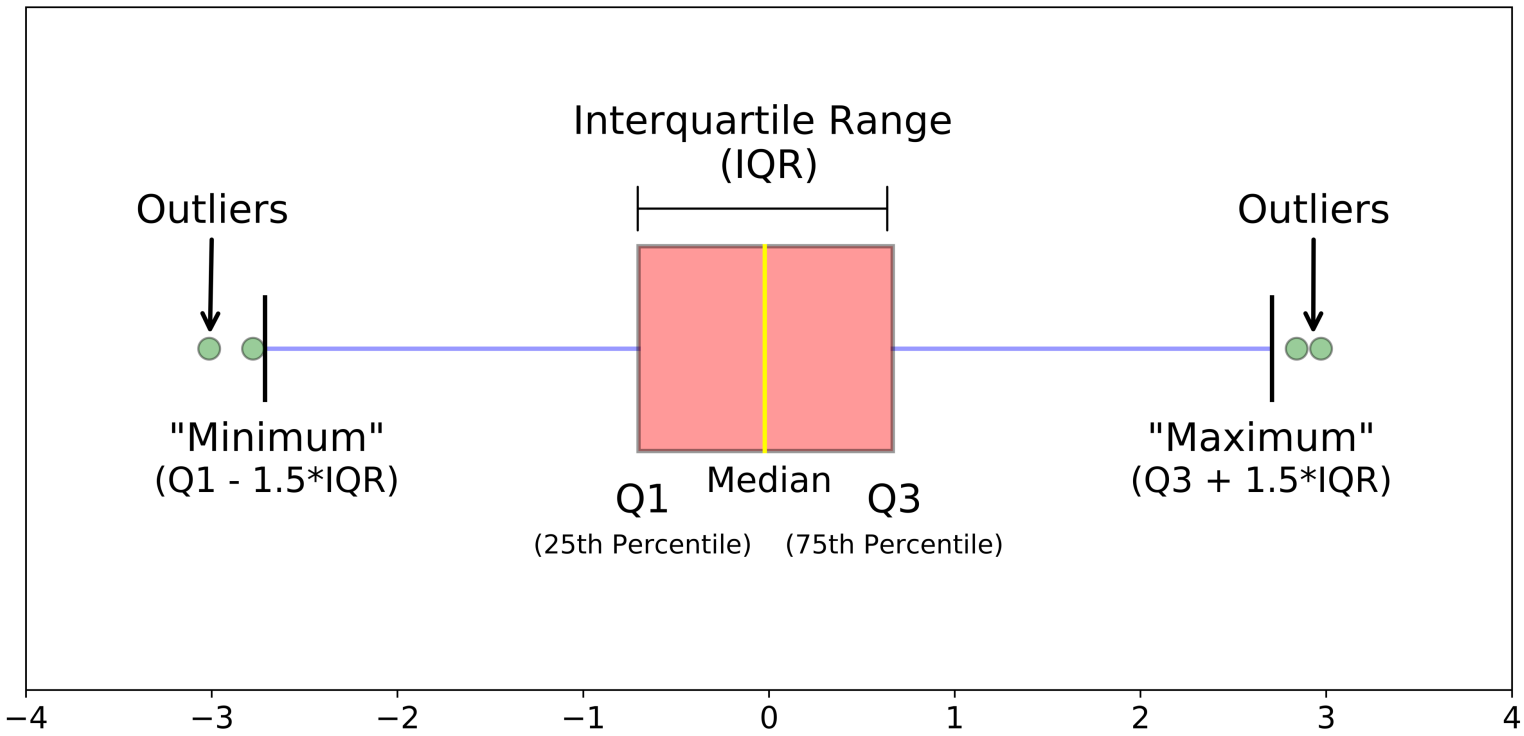
Statistike

- Mod (eng. mode) Vrednost koja se najčešće pojavljuje u skupu podataka.
- Medijana (eng. median). Vrednost koja deli slučajeve na pola nakon sortiranja. Ako postoji paran broj slučajeva, medijan je prosek dva srednja slučaja kada se sortiraju po rastućem ili opadajućem redosledu.

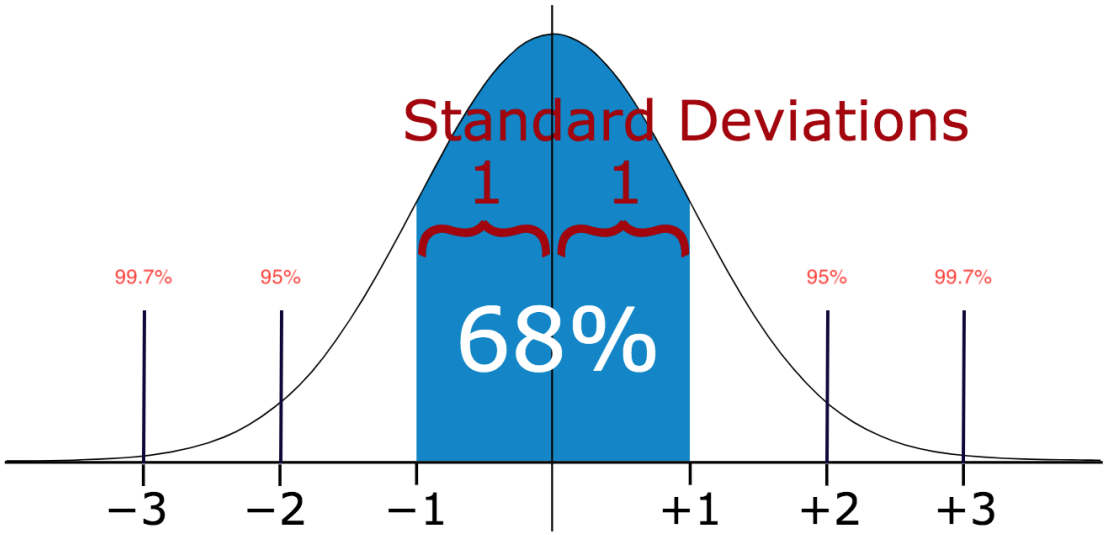
Statistike

- Postotna vrednost ili percentil za neki izabrani broj p definiše se poštujući uslov da je barem $p\%$ vrednosti u skupu manje ili jednako toj vrednosti
  - 25. percentil je poznat i kao prvi kvartil (Q_1)
  - 50. percentil je medijana ili drugi kvartil (Q_2)
  - 75. percentil je poznat i kao treći kvartil (Q_3)
  - interkvartilni raspon ($Q_3 - Q_1$)

Određivanje elemenata van granica korišćenjem percentila



Određivanje elemenata van granica metodom standardne devijacije



Obrada elemenata van granica i ekstremnih vrednosti

- *Coerce* - zamena elemenata van granica i ekstremnih vrednosti sa najbližom vrednošću koja se ne smatra elementom van granica. Npr, ako je element van granice definisan kao vrednost iznad ili ispod 3 standardne devijacije, onda bi svi elementi van granica bili zamenjeni najvećom ili najmanjom vrednošću unutar ovog opsega.
- *Discard* - odbacuju se slogovi sa elementima van granica u tom atributu

Obrada elemenata van granica i ekstremnih vrednosti

- *Nullify* - zamena elemenata van granica i ekstremnih vrednosti sa null ili sistemski nedostajućom vrednošću
- *Coerce outliers / discard extremes*
- *Coerce outliers / nullify extremes*

Generate > Outlier & Extreme SuperNode

Obrada *nedostajućih* vrednosti

- Null ili sistemski nedostajuće vrednosti - označene kao $null$
- prazne niske i beline - regularne vrednosti, mogu se definisati kao *blanko* vrednosti (čvorovi Database, Types...)
- *Blanko* ili korisnički definisane nedostajuće vrednosti u čvorovima Database, Types... (npr. 99 ili -1)

Obrada *nedostajućih* vrednosti - metod

- *Fixed* - zamena koristeći zadatu vrednost koja je rezultat izabrane statistike ili zadata konstanta
- *Random* - zamena izborom slučajne vrednosti
- *Expression* - zamena rezultatom zadatog izraza
- *Algorithm* - zamena korišćenjem vrednosti predviđene modelom dobijenog algoritmom C&RT.

Generate > Missing Values SuperNode

Outline

1 Istraživanje podataka

2 Skupovi podataka i atributi

3 Šum i elementi van granice

4 IBM SPSS Modeler

5 Učitavanje podataka

6 Zadatak

Primenom IBM SPSS Modeler uraditi:

- Učitati skup podataka *skupV1.xlxs* pomoću čvora Excel.
- Definisati da se atribut *id* ne koristi.
- Za aribut *age* promeniti gornju granicu u intervalu mogućih vrednosti na 100.
- U slogovima koji nemaju definisano ime, tj. uneti su prazne niske, zameniti praznu nisku sa jednom od niski iz liste (aa, bb, cc). Koristiti funkciju *oneof*.

Primenom IBM SPSS Modeler uraditi:

- Za atribut *gender*
  - prazne niske zameniti sa vrednošću *M*
  - postaviti da je binarnog tipa
  - postaviti da je vrednost *F* tačna vrednost
- Eliminisati slogove koji sadrže ekstremne vrednosti određene metodom sa kvartilima.
- Rezultat sačuvati u datoteci output.csv.