Normalizacija podataka stepenim transformacijama *Box-Cox i Yeo-Johnson transformacije*

Sara Živković

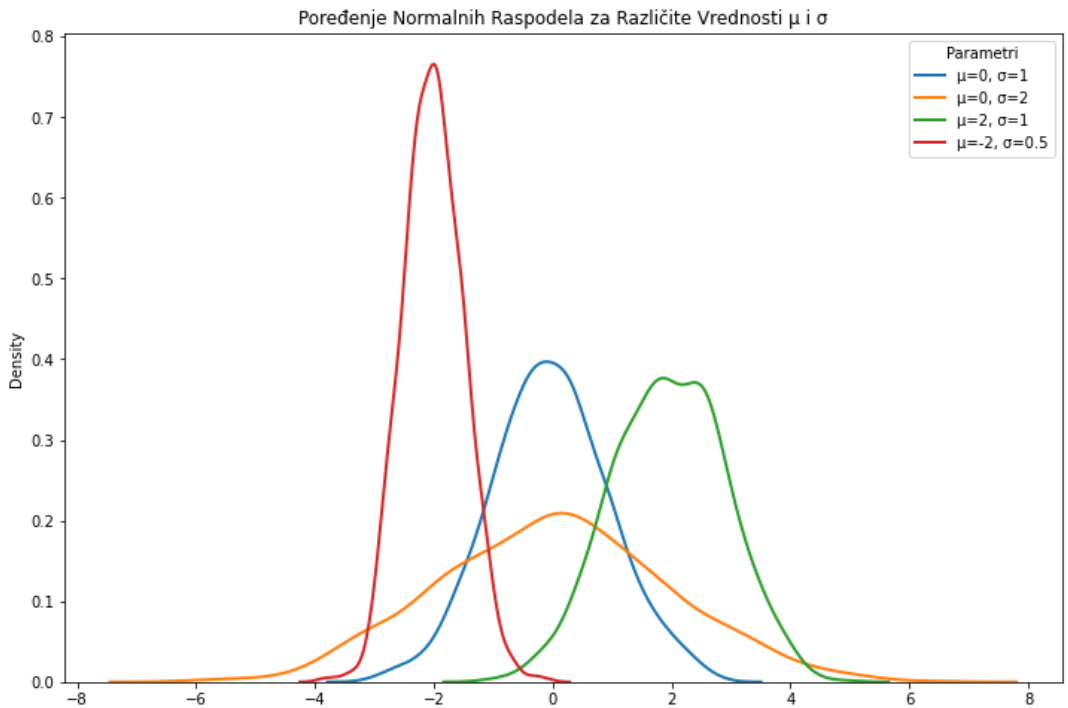
Matematički fakultet, Univerzitet u Beogradu

22. septembar 2024.

Sadržaj

Normalna raspodela

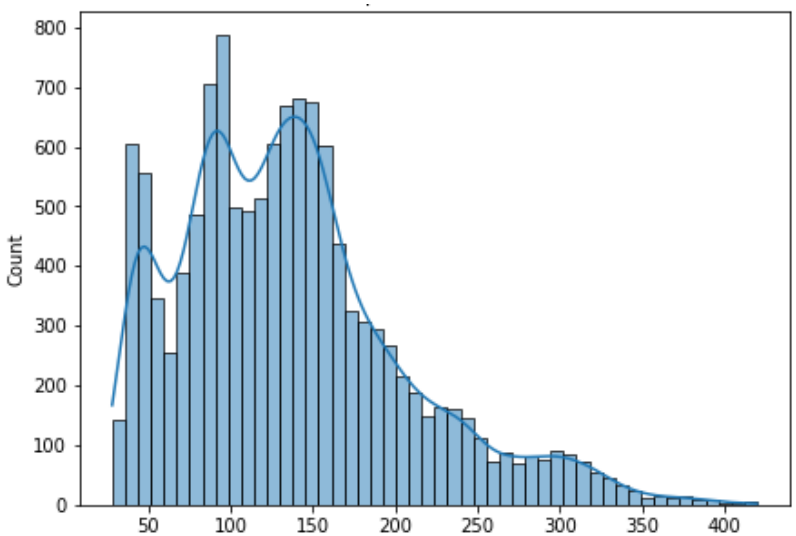
- Oblik zvonaste krive, simetrična oko svoje srednje vrednosti.
- Podaci su simetrično raspoređeni.
- Najveći broj podataka gravitira ka sredini raspodele.
- Ekstremne vrednosti imaju nisku, ali nenultu verovatnoću pojavljivanja.
- Srednja vrednost i medijana su iste vrednosti i nalaze se u centru krive.

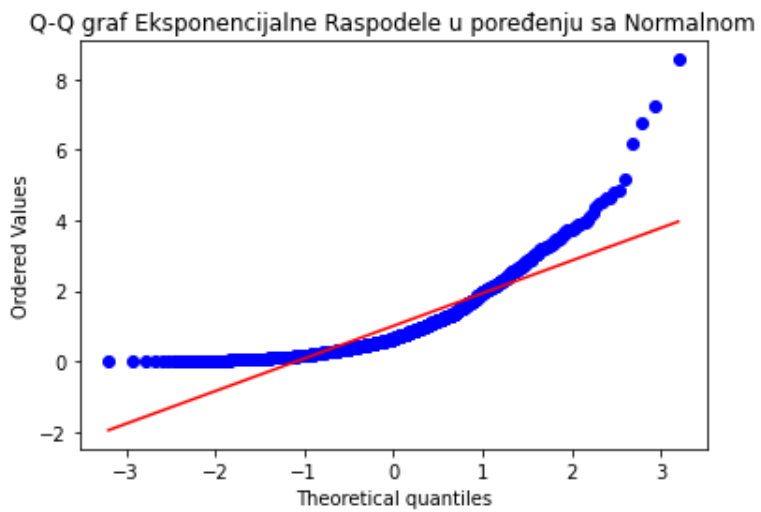


Kako proveriti raspodelu našeg uzorka?

⟶ Testovi normalnosti

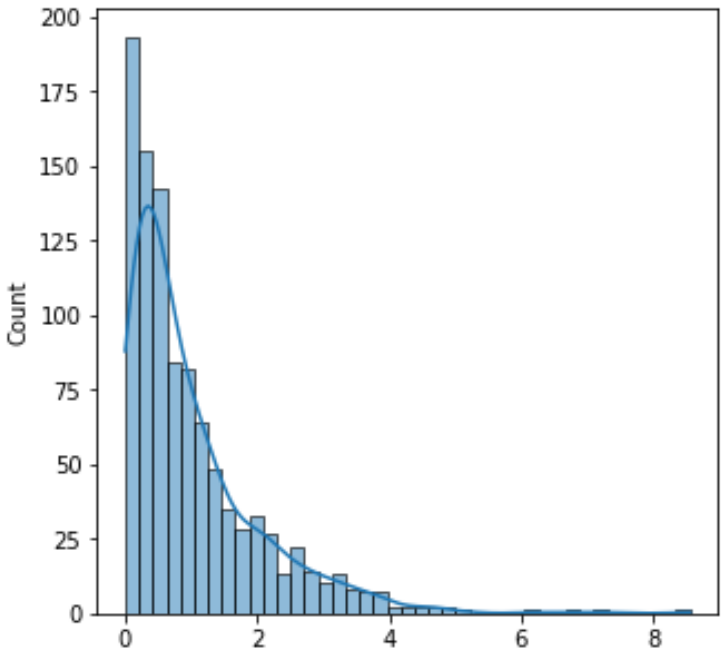
- Grafička reprezentacija raspodele
  - Grafik raspodele
  - Histogram
  - QQ grafik
- Metrike
  - Asimetrija ('Skewness')
  - Mera špicastosti ('Kurtosis')
- Statistički testovi normalnosti





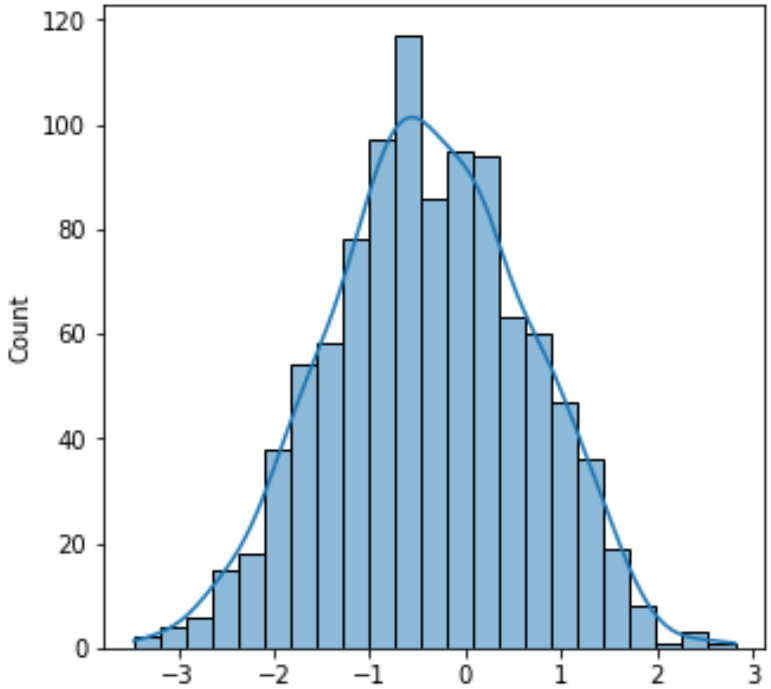
Šta ako podaci ne prate normalnu raspodelu? - Koristimo normalizaciju!

Primer: Generišemo skup slučajnih 1000 vrednosti iz eksponencijalne raspodele.



Pre normalizacije

Skew: 2.053



Posle normalizacije

Skew: 0.242

Stepene transformacije – put do normalizacije

- Logaritamska $y' = log(y)$
- Inverzna $y' = y^{-1}$
- Korene $y' = y^{1/2}, y' = y^{1/3}$
- **Box-Cox**

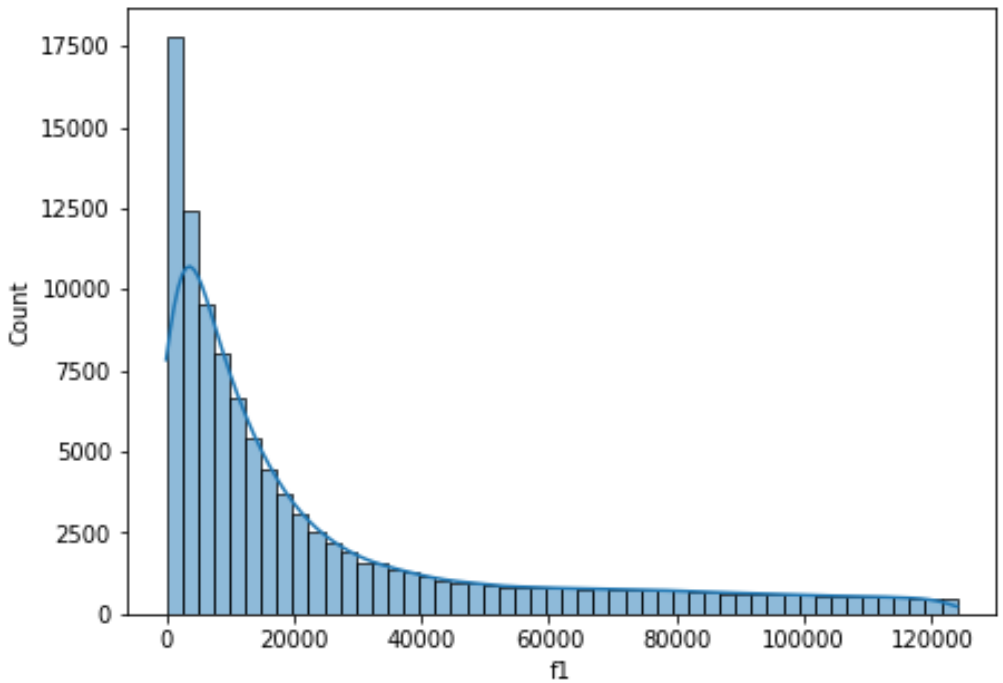
$$y'(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{ako } \lambda \neq 0, \\ \ln(y) & \text{ako } \lambda = 0. \end{cases}$$

- **Yeo-Johnson**

$$y'(\lambda) = \begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{ako } \lambda \neq 0 \text{ i } y \geq 0, \\ \frac{-(|y|+1)^{2-\lambda} + 1}{2-\lambda} & \text{ako } \lambda \neq 2 \text{ i } y < 0, \\ \ln(y+1) & \text{ako } \lambda = 0 \text{ i } y \geq 0, \\ -\ln(|y|+1) & \text{ako } \lambda = 2 \text{ i } y < 0. \end{cases}$$

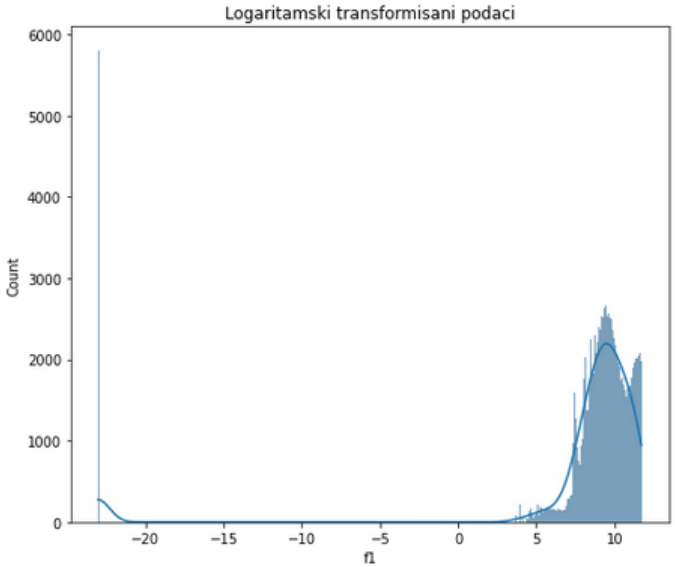
Primer

Uzorak koji koristimo: pregledi na web sajtu po minutu, u vremenskom periodu od 3 meseca. [link ka podacima]

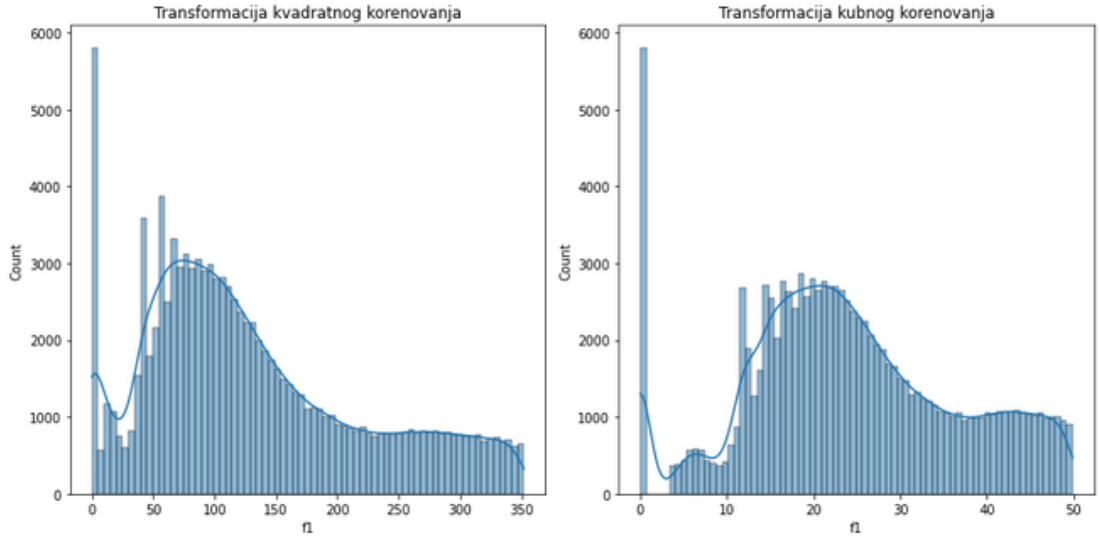


Slika: Histogram raspodele uzorka

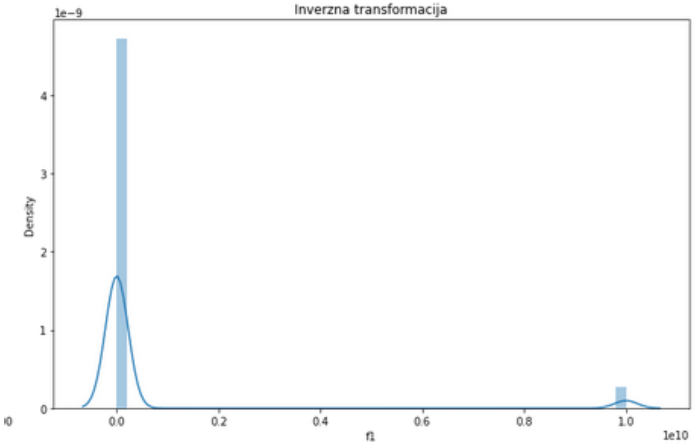
Zaključak: Uzorak ne prati normalnu raspodelu.



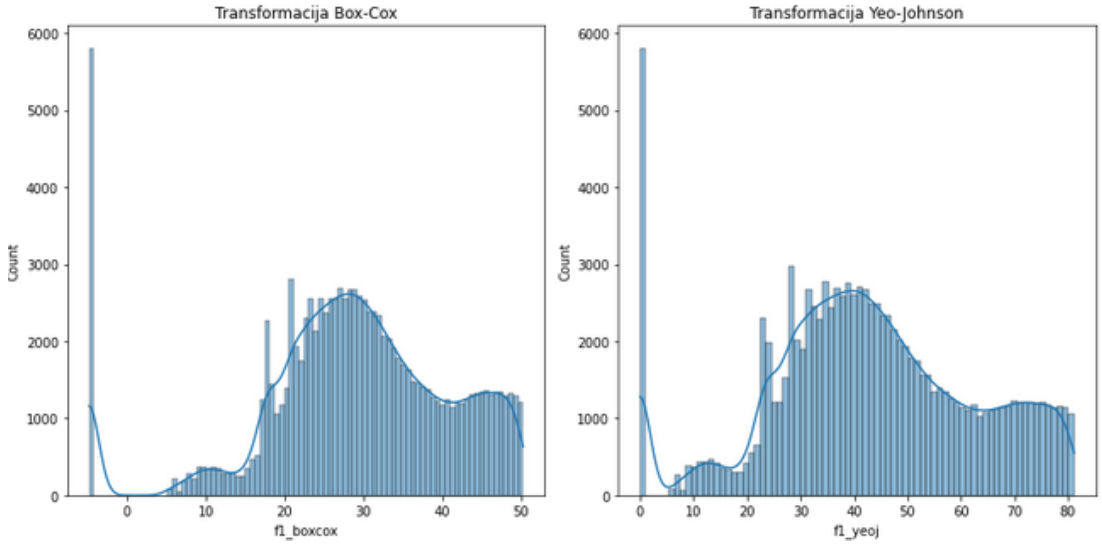
Logaritamska transformacija



Korenske transformacije



Inverzna transformacija



Box-Cox i Yeo-Johnson

Pored grafičkog prikaza, računali smo i razliku srednje vrednosti i medijane svake transformacije, kao i njenu asimetriju.
Rezultati:

Transformacija	**Razlika**	**Asimetrija**
Original	13342.37	1.54
Logaritamska	1.73	-3.72
Inverzna	13341.94	3.94
Kvadratni koren	21.59	0.70
Kubni koren	1.33	0.14
Box-Cox	**0.29**	1.54
Yeo-Johnson	**1.05**	1.54

Zaključak: Vidimo da su podaci najbliži normalnoj raspodeli primenom Box-Cox i Yeo-Johnson transformacija, ali da najpribližniju asimetriju normalnoj ima transformacija kubnim korenom.

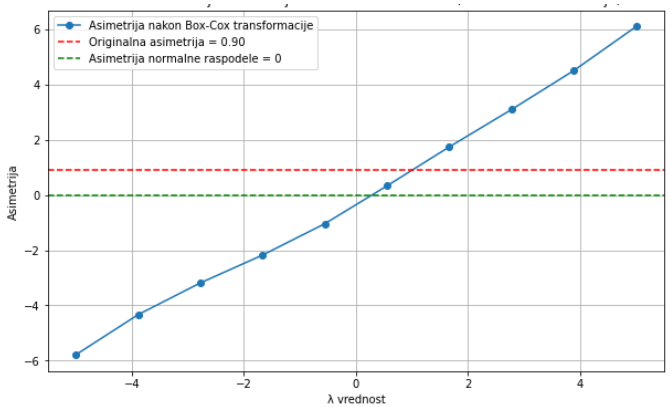
Box-Cox transformacija

- Radi samo sa pozitivnim podacima.
- Stepena transformacija koja koristi parametar λ da prilagodi raspodelu podataka.
  - $\lambda = 0.50$: transformacija pomoću kvadratnog korena
  - $\lambda =$ -1.00: transformacija pomoću recipročne (inverzne) vrednosti
  - $\lambda = 0.00$: transformacija pomoću prirodnog logaritma

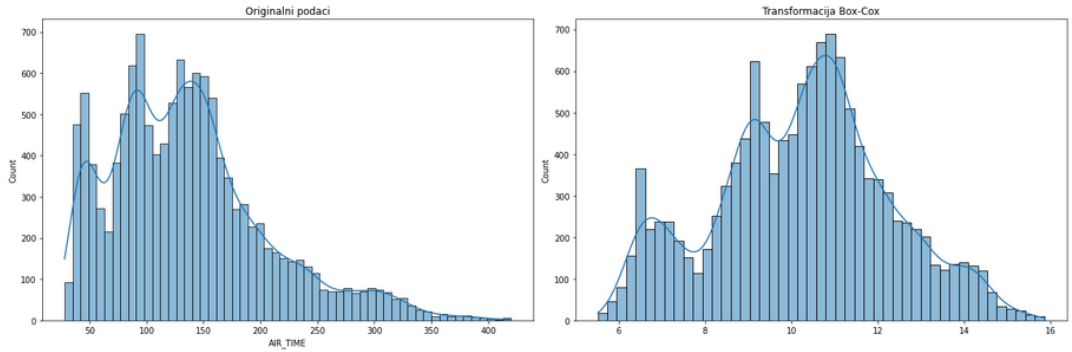
- Najčešće za sprovođenje transformacije koristi funkcija **boxcox** iz biblioteke **scipy.stats** koja omogućava automatsko određivanje optimalne vrednosti λ.

Primer

Uzorak koji koristimo: link ka skupu podataka



Slika: Optimalna vrednost λ prema boxcox funkciji: 0.28



Slika: Promena raspodele podataka nakon primene Box-Cox

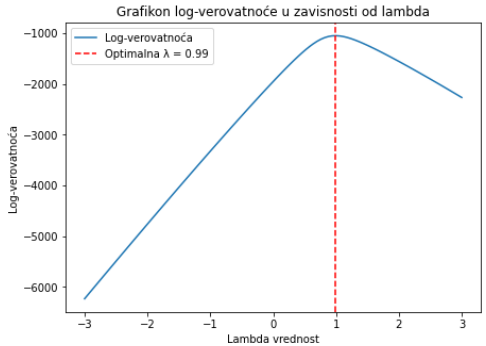
Yeo-Johnson transformacija

- Za podatke koji sadrže i pozitivne i negativne vrednosti.
- Parametar λ kontroliše oblik transformacije.

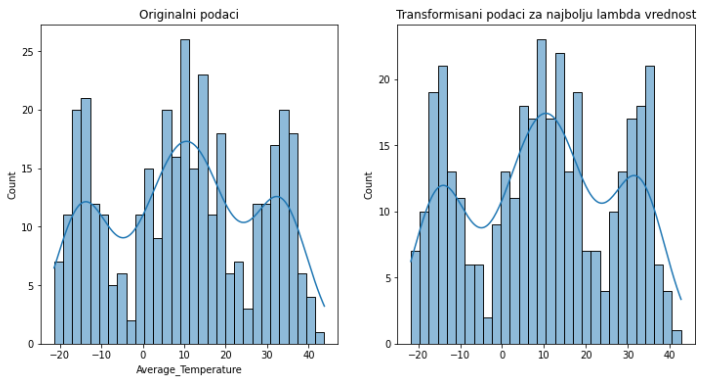
- Za sprovođenje transformacije koristi se uglavnom funkcija **yeojohnson** iz biblioteke **scipy.stats** koja omogućava automatsko određivanje optimalne vrednosti λ

Primer

Uzorak koji koristimo: Prosečne dnevne temperature u toku godine. [link ka skupu podataka]



Slika: Optimalna vrednost λ računanjem log-verovatnoće: 0.99



Slika: Promena raspodele podataka nakon primene Yeo-Johnson

Izazovi u radu

- Veza sa statističkom analizom i metodama.
- Pronaći odgovarajuće, reprezentativne skupove podataka.
- Potvrditi da je primenjena transformacija ispravno normalizovala podatke.
- Odlučiti se za najbolji pristup prikazivanja tranformacija.

Hvala na pažnji :)