nedeljainformatike v5.0

Istraživanje podataka: Kako od (male) šume videti drvo

Luka Jovičić

Matematička gimnazija

NEDELJA INFORMATIKE v5.0

21. decembar 2018.

Struganje

Struganje

Struganje – priča s tužnim krajem

- Aaron Swartz

Struganje – priča s tužnim krajem

- Aaron Swartz
- Javni podaci – (načelno) okej
- Potencijalno privatni podaci – – izbegavati

Kako stružemo

- RegEx[1]
- DOM
- Vizuelno/OCR
- ML

[1] https://stackoverflow.com/a/1732454/2363015

Kako stružemo

- RegEx[1]
- DOM
- Vizuelno/OCR
- ML

[1] https://stackoverflow.com/a/1732454/2363015

Kako stružemo

- RegEx[1]
- DOM
- Vizuelno/OCR
- ML

[1] https://stackoverflow.com/a/1732454/2363015

Kako stružemo

- RegEx[1]
- DOM
- Vizuelno/OCR
- ML

[1] https://stackoverflow.com/a/1732454/2363015

Kako stružemo

- RegEx[1]
- **DOM**
- Vizuelno/OCR
- ML

[1] https://stackoverflow.com/a/1732454/2363015

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

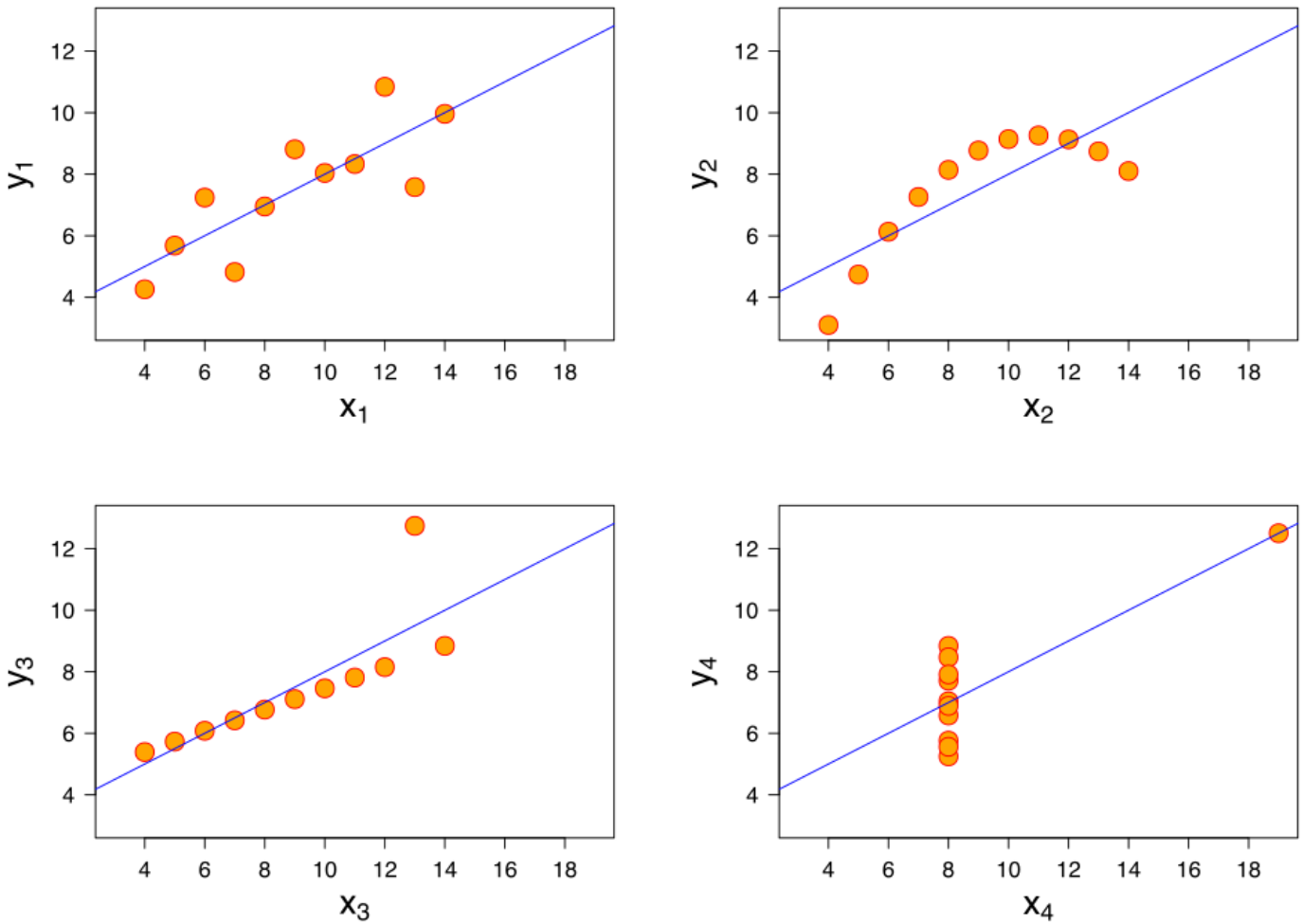
- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

Kako protiv strugača

- Nikako!
- Zahtevati login
- Rate limiting / captcha
- Honeypot-ovi
- Dinamički generisati HTML
- Koristiti nečitljive formate
- Lepo ih zamolite i nadajte se najboljem

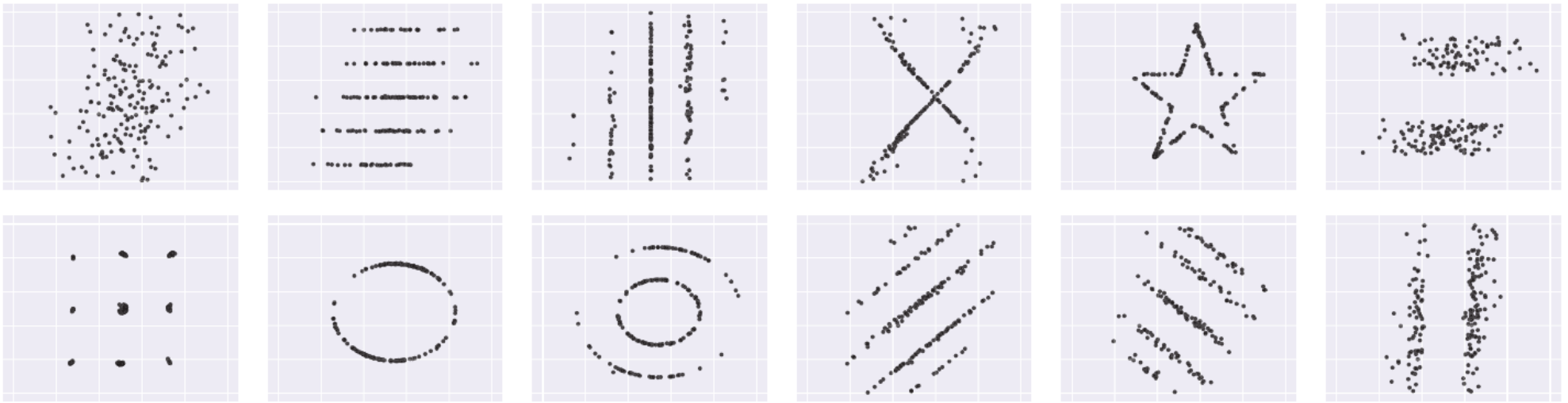
Anscombe's quartet

Anscombe's quartet



https://commons.wikimedia.org/w/index.php?curid=9838454

Same stats, different graphs



Matejka, J., & Fitzmaurice, G. (2017, May).

Istraživanje?

- John Tukey – Exploratory data analysis (EDA)
- Zahteva kreativnost i mnoštvo pitanja
- Varijacije
- Kovarijacije

Istraživanje?

- John Tukey – Exploratory data analysis (EDA)
- Zahteva kreativnost i mnoštvo pitanja
- Varijacije
- Kovarijacije

Istraživanje?

- John Tukey – Exploratory data analysis (EDA)
- Zahteva kreativnost i mnoštvo pitanja
- Varijacije
- Kovarijacije

Uzorak?

- (Nesavršen) deo populacije
- Istraživanjem opisujemo podatke
- Istraživanje $\not\Rightarrow$ predikcija
- Bavićemo se malim podacima (small data)

Uzorak?

- (Nesavršen) deo populacije
- Istraživanjem opisujemo podatke
- Istraživanje $\not\Rightarrow$ predikcija
- Bavićemo se malim podacima (small data)

Uzorak?

- (Nesavršen) deo populacije
- Istraživanjem opisujemo podatke
- Istraživanje $\not\Rightarrow$ predikcija
- Bavićemo se malim podacima (small data)

Uzorak?

- (Nesavršen) deo populacije
- Istraživanjem opisujemo podatke
- Istraživanje $\not\Rightarrow$ predikcija
- Bavićemo se malim podacima (small data)

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

Osnovni koraci

- Uvoz (import)
- Čišćenje (wrangling)
  - Otklanjanje grešaka
  - Tumačenje štrčaka (outliers)
  - Oblikovanje (reshaping)
- Vizuelizacija
- Transformacija
- Modeliranje
- Predstavljanje

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda. . .
- . . . ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda. . .
- . . . ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda...
- ...ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda...
- ...ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda...
- ...ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda...
- ...ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

R u 5 minuta

- Pravljen za matematičare $\Longrightarrow$ čudnovat jezik
- Vektor kao primarni tip podataka
- Funkcije imaju imenovane parametre
- Funkcije su građani prvog reda...
- ...ali postoje i objekti, na tri različita načina
- Često radimo sa data frame-ovima, a. k. a. nabudženim matricama
- `tidyverse`

Tibble

- *Slajdove odavde pa nadalje planiram da pređem kroz* `https://github.com/luka-j/csw5-eda`
- Tibble = fensi data frame
- `read_csv` → `tibble`
- Kolone – varijable (numeričke, kategoričke, tekstualne)
- Redovi – opservacije
- `$` za pristup kolonama (slično `.` u normalnim jezicima)

Tibble

- *Slajdove odavde pa nadalje planiram da pređem kroz* `https://github.com/luka-j/csw5-eda`
- Tibble = fensi data frame
- `read_csv` → `tibble`
- Kolone – varijable (numeričke, kategoričke, tekstualne)
- Redovi – opservacije
- `$` za pristup kolonama (slično `.` u normalnim jezicima)

Tibble

- *Slajdove odavde pa nadalje planiram da pređem kroz* `https://github.com/luka-j/csw5-eda`
- Tibble = fensi data frame
- `read_csv` $\rightarrow$ `tibble`
- Kolone – varijable (numeričke, kategoričke, tekstualne)
- Redovi – opservacije
- `$` za pristup kolonama (slično `.` u normalnim jezicima)

Tibble

- *Slajdove odavde pa nadalje planiram da pređem kroz* `https://github.com/luka-j/csw5-eda`
- Tibble = fensi data frame
- `read_csv` $\rightarrow$ `tibble`
- Kolone – varijable (numeričke, kategoričke, tekstualne)
- Redovi – opservacije
- `$` za pristup kolonama (slično `.` u normalnim jezicima)

Tibble

- *Slajdove odavde pa nadalje planiram da pređem kroz* `https://github.com/luka-j/csw5-eda`
- Tibble = fensi data frame
- `read_csv` $\rightarrow$ `tibble`
- Kolone – varijable (numeričke, kategoričke, tekstualne)
- Redovi – opservacije
- `$` za pristup kolonama (slično `.` u normalnim jezicima)

Plotovanje

- `ggplot2`
- "Ređamo" slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Plotovanje

- `ggplot2`
- “Ređamo” slojeve plota, spajamo sa `+`
- `geom`, `stat` i `scale`
- `geom_point` – scatter plot
- Overplotting
  - Transparentnost
  - Jitter
  - Kernel density estimate
  - 2d count (tabulacija)
- Više od dve dimenzije?
  - color
  - facet

Join, filter, transform

- `dplyr`
- `%>%` – pipe operator
  - `a %>% b(c)` $\longleftrightarrow$ `b(a,c)`
- Join spaja tabele koje imaju nešto zajedničko
  - `left`, `right`, `inner`, `full`, `semi`, `anti`
- `data <- data %>% filter(!štrčak)`
- Sve operacije sa kolonama se izvršavaju nad svim redovima

Join, filter, transform

- `dplyr`
- `%>%` – pipe operator
  - `a %>% b(c)` $\iff$ `b(a,c)`
- Join spaja tabele koje imaju nešto zajedničko
  - `left`, `right`, `inner`, `full`, `semi`, `anti`
- `data <- data %>% filter(!štrčak)`
- Sve operacije sa kolonama se izvršavaju nad svim redovima

Join, filter, transform

- `dplyr`
- `%>%` – pipe operator
  - `a %>% b(c)` $\iff$ `b(a,c)`
- Join spaja tabele koje imaju nešto zajedničko
  - `left`, `right`, `inner`, `full`, `semi`, `anti`
- `data <- data %>% filter(!štrčak)`
- Sve operacije sa kolonama se izvršavaju nad svim redovima

Join, filter, transform

- `dplyr`
- `%>%` – pipe operator
  - `a %>% b(c)` $\iff$ `b(a,c)`
- Join spaja tabele koje imaju nešto zajedničko
  - `left`, `right`, `inner`, `full`, `semi`, `anti`
- `data <- data %>% filter(!štrčak)`
- Sve operacije sa kolonama se izvršavaju nad svim redovima

Join, filter, transform

- `dplyr`
- `%>%` – pipe operator
  - `a %>% b(c)` $\iff$ `b(a,c)`
- Join spaja tabele koje imaju nešto zajedničko
  - `left`, `right`, `inner`, `full`, `semi`, `anti`
- `data <- data %>% filter(!štrčak)`
- Sve operacije sa kolonama se izvršavaju nad svim redovima

Linearni modeli

- George Box: “All models are wrong, but some are useful”
- `modelr`
- Težimo jednostavnim (i razumljivim!) modelima
- Cilj 1 – što manje reziduala
- Cilj 2 – haotična distribucija reziduala

Linearni modeli

- George Box: “All models are wrong, but some are useful”
- `modelr`
- Težimo jednostavnim (i razumljivim!) modelima
- Cilj 1 – što manje reziduala
- Cilj 2 – haotična distribucija reziduala

Linearni modeli

- George Box: “All models are wrong, but some are useful”
- `modelr`
- Težimo jednostavnim (i razumljivim!) modelima
- Cilj 1 – što manje reziduala
- Cilj 2 – haotična distribucija reziduala

Linearni modeli

- George Box: “All models are wrong, but some are useful”
- `modelr`
- Težimo jednostavnim (i razumljivim!) modelima
- Cilj 1 – što manje reziduala
- Cilj 2 – haotična distribucija reziduala

Linearni modeli

- George Box: “All models are wrong, but some are useful”
- `modelr`
- Težimo jednostavnim (i razumljivim!) modelima
- Cilj 1 – što manje reziduala
- Cilj 2 – haotična distribucija reziduala

Bonus: stringovi (i zašto ne)

- `stringr`
- Tekstualni podaci su (uglavnom) nepogodni za rad
- Regularni izrazi (akcenat na regularni)
- Srpski jezik: gomila prefiksa, sufiksa i glasovnih promena

Bonus: stringovi (i zašto ne)

- `stringr`
- Tekstualni podaci su (uglavnom) nepogodni za rad
- Regularni izrazi (akcenat na regularni)
- Srpski jezik: gomila prefiksa, sufiksa i glasovnih promena

Bonus: stringovi (i zašto ne)

- `stringr`
- Tekstualni podaci su (uglavnom) nepogodni za rad
- Regularni izrazi (akcenat na regularni)
- Srpski jezik: gomila prefiksa, sufiksa i glasovnih promena

Bonus: stringovi (i zašto ne)

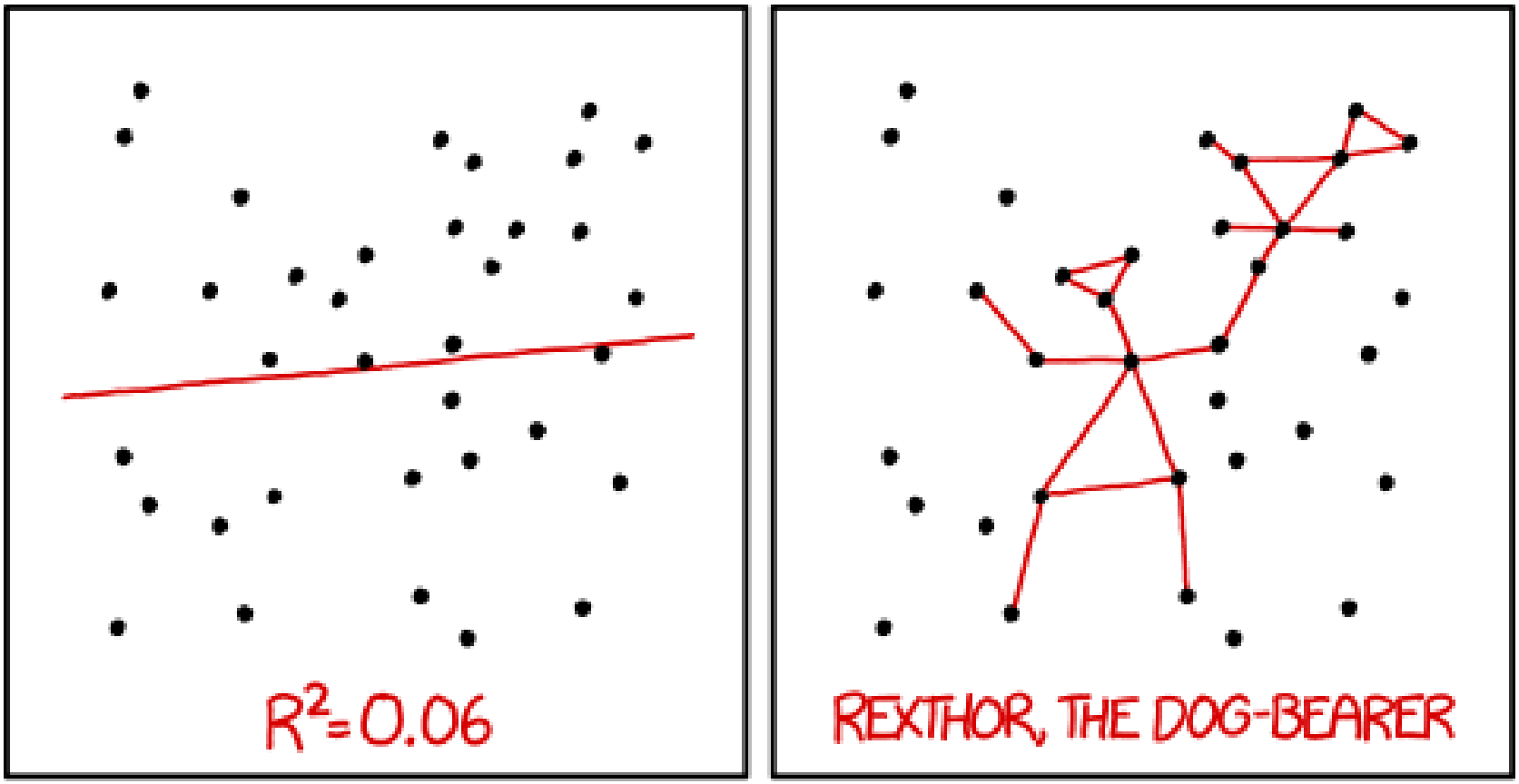
- `stringr`
- Tekstualni podaci su (uglavnom) nepogodni za rad
- Regularni izrazi (akcenat na regularni)
- Srpski jezik: gomila prefiksa, sufiksa i glasovnih promena

Zaključak

- Vizuelizacije su kul
- Predikcija je kul, ali ne treba zanemariti razumevanje

Zaključak

- Vizuelizacije su kul
- Predikcija je kul, ali ne treba zanemariti razumevanje



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Bonus: `https://xkcd.com/2048`