Datu analīze ar Python

Metodes, piemēri, vizualizācijas

2023. gada 19. janvāris

Jānis Lazovskis, RTU Riga Business School

SDSK7061 : Jaunākās tendences sociālo zinātņu pētījumu metodēs

Ievads: personīgi

bakalaura grāds:	2009 - 2013: University of Waterloo, Canada
maģistra grāds:	2013 - 2014: University of Waterloo, Canada
doktora grāds:	2014 - 2019: University of Illinois at Chicago, USA
pēcdoktorantūra:	2019 - 2020: University of Aberdeen, UK
lektors / viesdocents:	2020 - 2021: Latvijas Universitāte, LV
lektors / docents:	2020 - 2023: RTU Riga Business School, LV

Ievads: par šo nodarbību

Kodēšana:

- Nav obligāta, nebūs tikai Python
- Lietošu Google Colaboratory (colab.research.google.com) interaktīvo darba vidi

Valoda:

- Tehniskie termini bieži nāk no angļu valodas
- Latviskoti pēc iespējas

Jūs:

- Kāda pieredze?
- Kas sagādjis grūtības?
- Kas pietrūkst?
- Ko ļoti vēlos?

Datu analīzes pamati

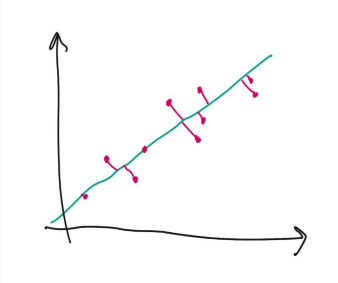
Motivācija: Partiju šķirotava

Motivācija: Partiju šķirotava

- Vairāku mēnešu Providus pētījums, LSM realizācija

- 11 partijām tiek doti 58 apgalvojumi, jāatbild no “pilnībā nepiekrītu” līdz “pilnībā piekrītu” (5 iespējas)

- Kā es to uztvēru:
  - 11 punkti 58 dimensiju telpā
  - Bet politikā nav tik daudz dimensiju - jābūt vienkāršākai telpai, kas visus viedokļus aptver



<- divās dimensijās

ir (gandrīz) tas pats kas

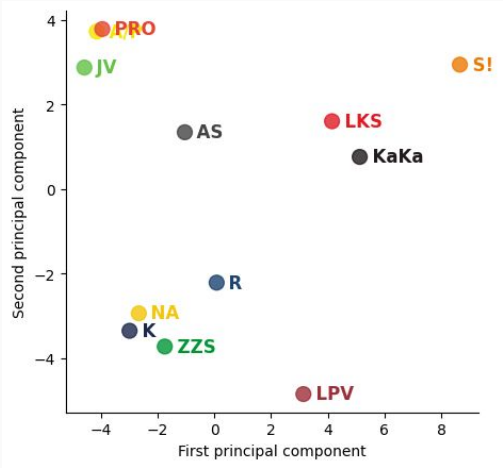
vienā dimensijā ->

Partiju šķirotava: Avoti un metode

- Šķirotava LSM: https://www.lsm.lv/velesanas2022/partiju-skirotava/
- Par projekta tapšanu: https://www.lsm.lv/raksts/arpus-etera/arpus-etera/par-projektu-partiju-skirotava-pirms-14-saeimas-velesanam.a471344/
- Secinājumi no Providus par šķirotavu: https://providus.lv/raksti/veletaju-skirotava-vai-partiju-atbalstitaju-viedokli-sakrit-ar-politisko-speku-domam-originals-no-lsm-lv/

---

1. Pārkopēt pirmkodu no katras partijas atbildēm
2. Atlasīt atbildes ar Python
3. Pārveidot atbildes par ciparіem un salikt tabulā
4. Atrast tuvāko plakni visiem punktiem, to parādīt



Partiju šķirotava: Datu skrāpēšana

* Python notebook

- Atbildes publicētas atsevišķi pēc partijām, daļēji stukturētā HTML kodā
- Providus vēlāk lietojamā veidā visas atbildes publicēja

```
<div class="party-answers">
  <div class="quiz-answers-list">
    <div class="item">
      <h4 class="question"> flex
        ::before
        Latvijā nav vajadzīgas "sarkanās līnijas" valdības koalīciju veidošanā
      </h4>
      <div class="answer"> flex
        <div class="party-answer">
          <strong>Partijas atbilde:</strong>
          pilnībā nepiekrītu
        </div>
      </div>
    </div>
```

```
pilnībā piekrītu:              2
daļēji piekrītu:               1
gan piekrītu, gan nepiekrītu:  0
daļēji nepiekrītu:            -1
pilnībā nepiekrītu:           -2
```

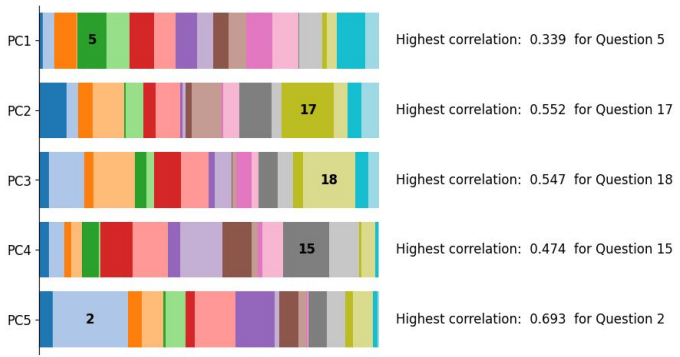
```
# Set up data frame dictionary
data_dict = {p:[] for p in party_list}

# Load files
for i,party in enumerate(party_list):
    print(party, flush=True)
    soup = BeautifulSoup(open(root+'atbildes-pirmais/'+party+'.html').read(), features='html.parser')
    answers = [ans.contents[1].strip() for ans in soup.find_all('div', {'class':['party-answer']})]
    data_dict[party] = [value_dict[ans] for ans in answers]

    # Save questions once
    if i==0:
        questions = [q.contents[0].strip() for q in soup.find_all('h4', {'class':['question']})]
        questions_dict = {j:questions[j] for j in range(len(questions))}
        with open(root+'jautajumi.pickle', 'wb') as handle:
            pickle.dump(questions_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)

# Create and save dataframe
df = pd.DataFrame.from_dict(data_dict)
df.to_pickle(root+'atbildes.pkl')
df.to_csv(root+'atbildes.csv')

print('Done', flush=True)
```

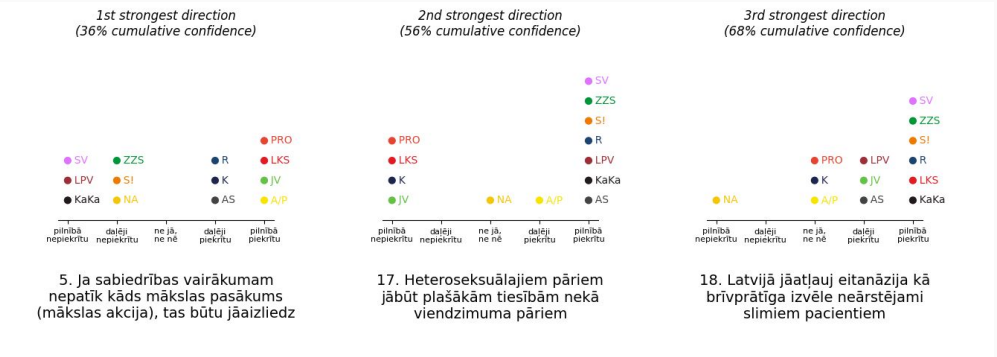
Galveno komponentu analīze GKA / Principal component analysis PCA

* Python notebook

1. Dati **jāpārbīda** uz vidējo katrā asī
2. Uztverot datus kā matricu, **jāaprēķina** "īpašvērtības" (jeb "eigenvalues")
3. **Projecē** visus datus uz plakni, ko veido pirmās divas īpašvērtības
4. **Izmērot** leņķus starp asīm un pirmām divām īpašvertībām, var aptuveni noteikt, kuras asis (koordinātes) ir noteicošās

---





5. Ja sabiedrības vairākumam nepatīk kāds mākslas pasākums (mākslas akcija), tas būtu jāaizliedz

17. Heteroseksuālajiem pāriem jābūt plašākām tiesībām nekā viendzimuma pāriem

18. Latvijā jāatļauj eitanāzija kā brīvprātīga izvēle neārstējami slimiem pacientiem

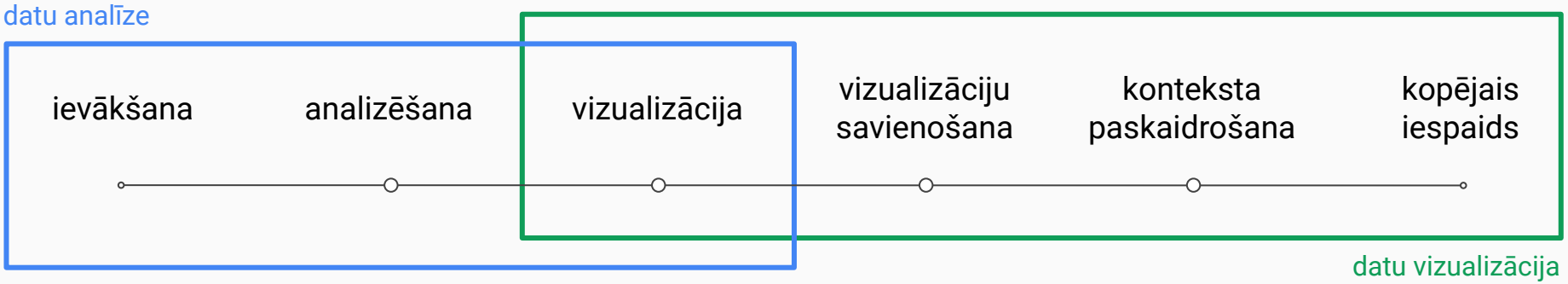
Partiju šķirotava: salīdzināt ar “Politisko tinderi”

- Factum pētījums, Ir realizācija
- 5 kandidātiem no 7 partijām tiek doti 18 jautājumi, jāatbild “jā” vai “nē”
- Atšķiras no “Partiju šķirotavas”, bet rezultāti ir salīdzināmi

Praktiskais darbs ar datiem

1. Datu dzīves cikls
1. Datu tipi
2. Python pakotnēs
3. Rīki ārpus Python

Datu dzīves cikls: pārskats



aptauju ievākšana

datu skrāpēšana

tīrīšana

atlasīšana

atlasīšana

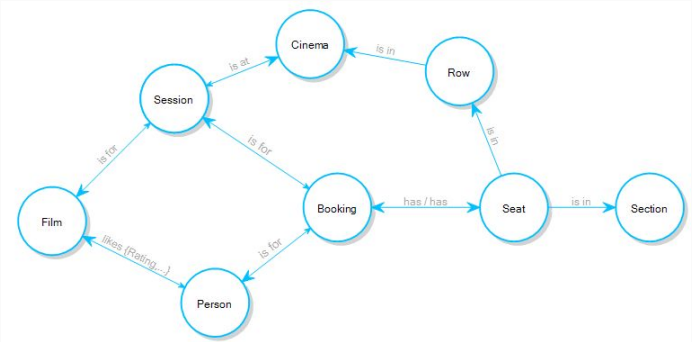
Datu tipi: strukturēti

* Python notebook

- **Tabulas** formā:
  - .xls, .xlsx (Microsoft Excel)
  - .csv, .tsv (vienkārši teksta faili)

	A	B	C	D	E	F
1	song_name	danceability	energy	key	loudness	mode
2	Mercury: Retrograde	0.831	0.814	2	-7.364	1
3	Pathology	0.719	0.493	8	-7.23	1
4	Symbiote	0.85	0.893	5	-4.783	1
5	ProductOfDrugs (Prod. The Virus and Antidote)	0.476	0.781	0	-4.71	1
6	Venom	0.798	0.624	2	-7.668	1
7	Gatteka	0.721	0.568	0	-11.295	1
8	kamikaze (+ pulse)	0.718	0.668	8	-4.162	1
9	T.R.U. (Totally Rotten Underground)	0.694	0.711	8	-5.525	1
10	I Put My Dick in Your Mental	0.774	0.751	1	-2.445	1
11	Andromeda	0.893	0.907	11	-10.406	1
12	BRAINFOOD	0.864	0.365	8	-10.219	1
13	Troll Under the Bridge	0.736	0.932	1	-3.726	1
14	1000 Rounds	0.825	0.761	8	-5.389	1
15	Sacrifice	0.767	0.576	10	-9.683	0
16	Backpack	0.765	0.726	5	-5.58	1
17	D(R)Own	0.617	0.541	6	-4.113	1
18	Okay,ButThisIsTheLastTime	0.755	0.298	1	-15.032	1
19	TakingOutTheTrash	0.814	0.575	11	-9.635	1

- **Nav tabulas** formā**:**
  - .json (vārdnīca - atslēga un vērtība)
  - koki un grafi (var atveidot ar vārdnīcas struktūru)

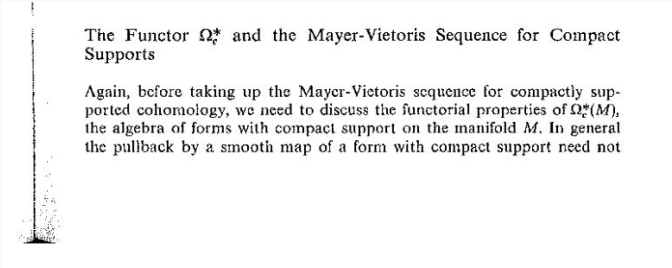



```
{
  "particles": {
    "number": {
      "value": 80,
      "density": {
        "enable": true,
        "value_area": 800
      }
    },
    "color": {
      "value": "#01b6ed"
    },
    "shape": {
      "type": "circle",
      "stroke": {
        "width": 0,
        "color": "#01b6ed"
      },
      "polygon": {
        "nb_sides": 10
      },
      "image": {
        "src": "img/github.svg",
        "width": 100,
        "height": 100
      }
```

Datu tipi: daļēji strukturēti, nestrukturēti

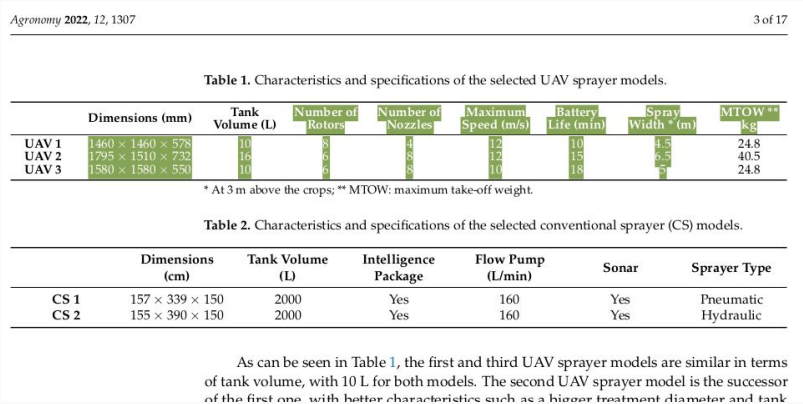
- **Daļēji strukturēti** (dati atpazīstami, bet to kārtība nav noteikta):
  - .html
  - .pdf (ar kopējamu tekstu)

- **Nestrukturēti** (dati neatpazīstami):
  - bildes
  - .pdf (ieskenēti, bez kopējama teksta)



The Functor Ω_c^* and the Mayer-Vietoris Sequence for Compact Supports

Again, before taking up the Mayer-Vietoris sequence for compactly supported cohomology, we need to discuss the functorial properties of $\Omega_c^*(M)$, the algebra of forms with compact support on the manifold M. In general the pullback by a smooth map of a form with compact support need not



Agronomy **2022**, *12*, 1307 3 of 17

Table 1. Characteristics and specifications of the selected UAV sprayer models.

	Dimensions (mm)	Tank Volume (L)	Number of Rotors	Number of Nozzles	Maximum Speed (m/s)	Battery Life (min)	Spray Width * (m)	MTOW ** kg
UAV 1	1460 × 1460 × 578	10	8	4	12	10	4.5	24.8
UAV 2	1795 × 1510 × 732	16	6	8	12	15	6.5	40.5
UAV 3	1580 × 1580 × 550	10	6	8	10	18	5	24.8

* At 3 m above the crops; ** MTOW: maximum take-off weight.

Table 2. Characteristics and specifications of the selected conventional sprayer (CS) models.

	Dimensions (cm)	Tank Volume (L)	Intelligence Package	Flow Pump (L/min)	Sonar	Sprayer Type
CS 1	157 × 339 × 150	2000	Yes	160	Yes	Pneumatic
CS 2	155 × 390 × 150	2000	Yes	160	Yes	Hydraulic

As can be seen in Table 1, the first and third UAV sprayer models are similar in terms of tank volume, with 10 L for both models. The second UAV sprayer model is the successor of the first one, with better characteristics such as a bigger treatment diameter and tank

```
<div class="party-answers">
  <div class="quiz-answers-list">
    <div class="item">
      <h4 class="question"> flex
        ::before
        Latvijā nav vajadzīgas "sarkanās līnijas" valdības koalīciju veidošanā
      </h4>
      <div class="answer"> flex
        <div class="party-answer">
          <strong>Partijas atbilde:</strong>
          pilnībā nepiekrītu
        </div>
      </div>
    </div>
```

Datu tipi: rīki informācijas izvilkšanai

* Python notebook

- **Daļēji strukturētiem failiem**
  - .html : skrāpēšana ar `BeautifulSoup` pakotni
  - .pdf : kopēšana uz Excel

- **Nestrukturētiem failiem**
  - .pdf un bildēm : dažādas lietotnes (`tesseract` priekš Linux, Microsoft Lens priekš Android, …)

Analīze caur vizualizāciju

1. Mērķauditorija un stāsts
2. Tālāki avoti

Mērķauditorija un stāsts: trīs jautājumi

Kuri ir mērķauditorijā?

- Kas viņiem ir svarīgs?
- Kas viņus motivē?
- Pie kā viņi ir pieraduši?

Ko es cenšos parādīt?

- Uz kādu rīcību es gribu mērķauditoriju virzīt?
- Ko es gribu, lai lasītajs secina?

Kā mani dati parādīs to, ko vēlos, lai parāda?

- Kāda vizualizācija ir vispiemērotākā?
- Kā uzsvērt maiņas / izņēmumu / kopskatu?

no C.N. Knaflic, *Storytelling with Data*

Tālāki avoti

- (grāmata) *Introduction to Linear Algebra*, Gilbert Strang
- (grāmata) *Storytelling with Data*, Cole Nussbaumer Knaflic
- (grāmata) *Building Science Graphics,* Jen Christiansen
- (žurnāls) *Nightingale*, the Data Visualization Society

Paldies par jūsu uzmanību