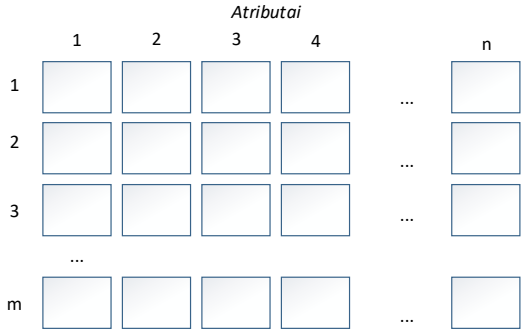
Laboratorinis darbas Nr.1. Duomenų apdorojimas ir analizė

1. Pasirinkti (susikurti) duomenų rinkinį[1,2,3], su kuriuo atliksite šį ei sekančius laboratorinius darbus. Jūsų pasirinkimą turi patvirtinti vienas iš laboratorinių darbų dėstytojų[4].

Duomenų rinkinio reikalavimai:

- Turi egzistuoti skaitinės (*integer* ir *real* tipo) ir /arba kategorinės reikšmės. Duomenų rinkinys kuriame yra tik kategorinio tipo atributai **yra netinkamas**.
- Duomenų rinkinyje įrašų (eilučių) *m* turi būti ne mažiau nei 500, t.y., $\infty > m \geq 500$ ir atributų *n* nemažiau nei 8 (stulpeliai) $\infty > n \geq 8$. Jeigu atributų *n* pasirinktame duomenų rinkinyje yra mažiau, privalote pridėti išvestinius (sukurtus) atributus (žr. pav. 1.)

Svarbu. Sekančios užduotys turi būti realizuotos programiškai naudojant *Python*.



pav. 1. Duomenų aibės grafinis atvaizdavimas

2. Atlikti duomenų rinkinio kokybės analizę (žr. 2 pav.). Kiekvienam **tolydinio** tipo atributui paskaičiuoti:
   - bendrą reikšmių skaičių,
   - trūkstamų reikšmių procentą,
- kardinalumą (***kardinalumas*** matematikoje yra aibės savybė , apibendrinanti baigtinės aibės narių kiekio sąvoką. Paprasčiau tariant kiek yra skirtingų atributo reikšmių. Pavyzdžiui lyties atributo kardinalumas lygus 2 – t.y., lytis gali tur4ti tik dvi reikšmes),
   - minimalią (*min*) ir maksimalią (*max*) reikšmes,
   - 1-ąją ir 3- ją kvartilius (žr. 2 paskaita, 37 skaidrę),
   - Vidurkį (žr. 2 paskaita, 36 skaidrę),
   - Medianą (žr. 2 paskaita, 36 skaidrę),
   - standartinį nuokrypį (žr. 2 paskaita, 36 skaidrę).
3. Kiekvienam **kategorinio** tipo atributui paskaičiuoti:
   - bendrą reikšmių skaičių,
   - trūkstamų reikšmių procentą,
   - kardinalumą,
   - modą (***moda*** - vadinama dažniausiai pasitaikanti imties reikšmė) (žr. 2 paskaita, 39 skaidrę),
   - modos dažnumo reikšmę (žr. 2 paskaita, 39 skaidrę),
   - modos procentinę reikšmę(žr. 2 paskaita, 39 skaidrę),

[1] Find Open Datasets and Machine Learning Projects | Kaggle
[2] https://archive.ics.uci.edu/ml/datasets.php
[3] https://vincentarelbundock.github.io/Rdatasets/datasets.html
[4] A.Tarasevičienė, G.Budnikas, A.Nečiūnas

- 2-ąją modą (žr. 2 paskaita, 39 skaidrę),
- 2-osios modos dažnumo reikšmę (žr. 2 paskaita, 39 skaidrę),
- 2-osios modos procentinę reikšmę (žr. 2 paskaita, 39 skaidrę).

Tolydinio tipo reikšmėms

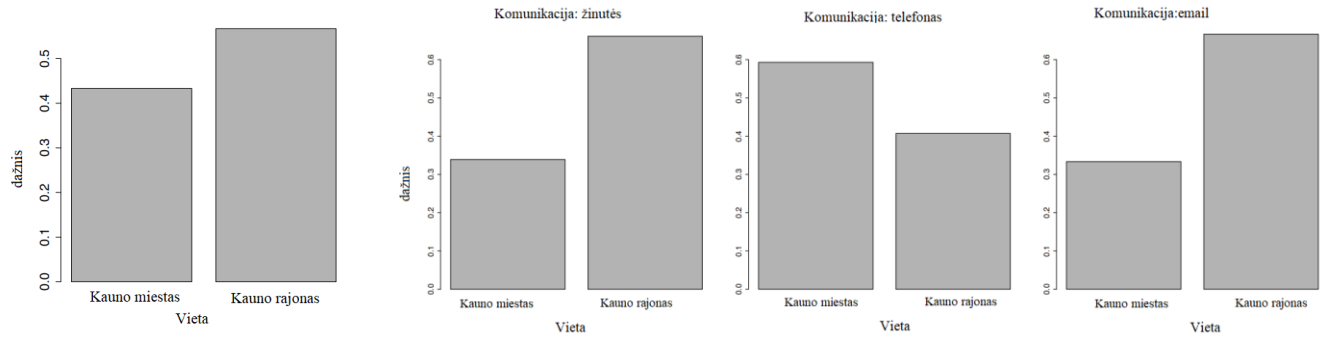
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-iasis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis

Kategorinio tipo reikšmėms

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %

pav. 2. Tolydinio ir kategorinio tipo duomenų analizės kokybės parametrų lentelės

4. Nupaišyti atributų histogramas (rekomenduotinas stulpelių skaičius randamas formule: $1 + 3.22 \cdot log_e^n$, kur n imties dydis). Ataskaitoje pateikti aprašymus, koks tai pasiskirstymas (pvz., *normalusis, vien(a)modalis, eksponentinis* ir t.t.) ir kokias išvadas pagal tai galima formuluoti (žr. 2 paskaita, 41-43 skaidres).
5. Identifikuoti duomenų kokybės problemas: trūkstamas reikšmes, kardinalumo problemas, triukšmus– ekstremalias reikšmes (angl. *outliers*) (žr. 2 paskaita, 46-53 skaidres).
6. Pateikti šių problemų sprendimo planą, kuris bus realizuotas programiškai (pvz., bus įtraukiamos trūkstamos kategorinio atributo reikšmės remiantis atributo moda įverčiu, ekstremalios reikšmės yra šalinamos ar koreguojamos).
7. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus:
   - **Tolydinio tipo atributams**: naudojant „scatter plot“ tipo diagramą (žr. 3 paskaita, 5 skaidrė) pateikti kelis (2-3) pavyzdžius su stipria tiesine atributų priklausomybe (tiesioginė arba atvirkštinė koreliacija) bei kelis pavyzdžius su tarpusavyje nekoreliuojančiais (silpnai koreliuojančiais) atributais. Pakomentuoti rezultatus.
   - Pateikti SPLOM diagramą (Scatter Plot Matrix) (žr. 3 paskaita, 6 skaidrė).
   - **Kategorinio tipo atributams**: naudojant „bar plot“ tipo diagramą pateikti keletą (2-3) atributų priklausomybės pavyzdžių ir pakomentuoti rezultatus (žr. 3 paskaita, 7-9 skaidres).
   - Pateikti keletą (2-3) histogramų (žr. 3 paskaita, 12-14 skaidres) ir „box plot“ diagramų pavyzdžių (žr. 3 paskaita, 15 skaidrę), vaizduojančių sąryšius tarp **kategorinio** (pavyzdys pateiktas pav.3) ir **tolydinio** tipo kintamųjų .
8. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą (žr. 3 paskaita, 24-34 skaidres). Rezultatus pakomentuoti.
9. Atlikti duomenų normalizaciją (rėžiai [0;1] arba [-1;1]) (žr. 3 paskaita, 35-37 skaidres).
10. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.



pav. 3. „Bar plot“ tipo diagrama atvaizduojanti: a) Vieno kategorinio tipo atributo“ Vieta“ histogramą; b) ir priklausomybę tarp dviejų kategorinio tipo atributų „Vieta“ ir „Komunikacija“.