მონაცემთა მეცნიერება

მანქანური სწავლების გამოყენებამდე

რევაზ ტატიშვილი

revaz.tatishvili@gmail.com

https://github.com/rtatishvili

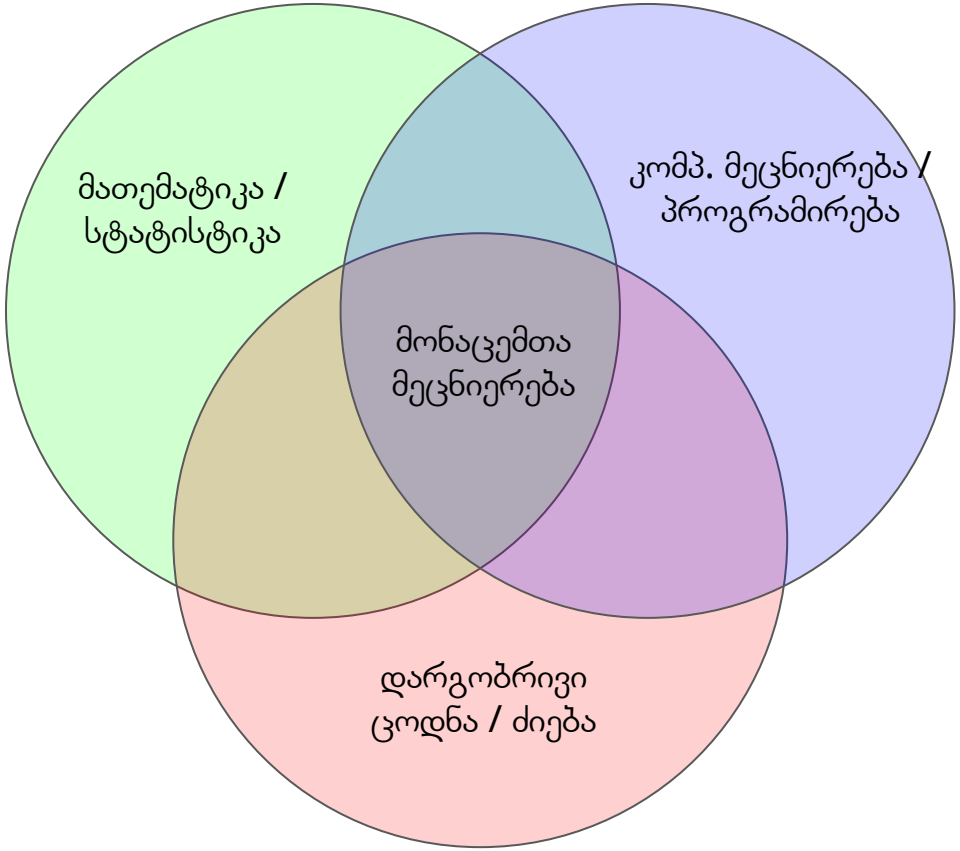
https://www.linkedin.com/in/revaz-tatishvili-79177a71/

დღის წესრიგი

- შესავალი
- მონაცემთა პირველადი გამოკვლევა
- მონაცემთა გაწმენდა და გამდიდრება
- სტატისტიკური ანალიზი
- “დემო”
- შეჯამება

შესავალი

მონაცემთა მეცნიერება არის მრავალდისციპლინური დარგი:



...

შესავალი

მონაცემთა მეცნიერება არის:

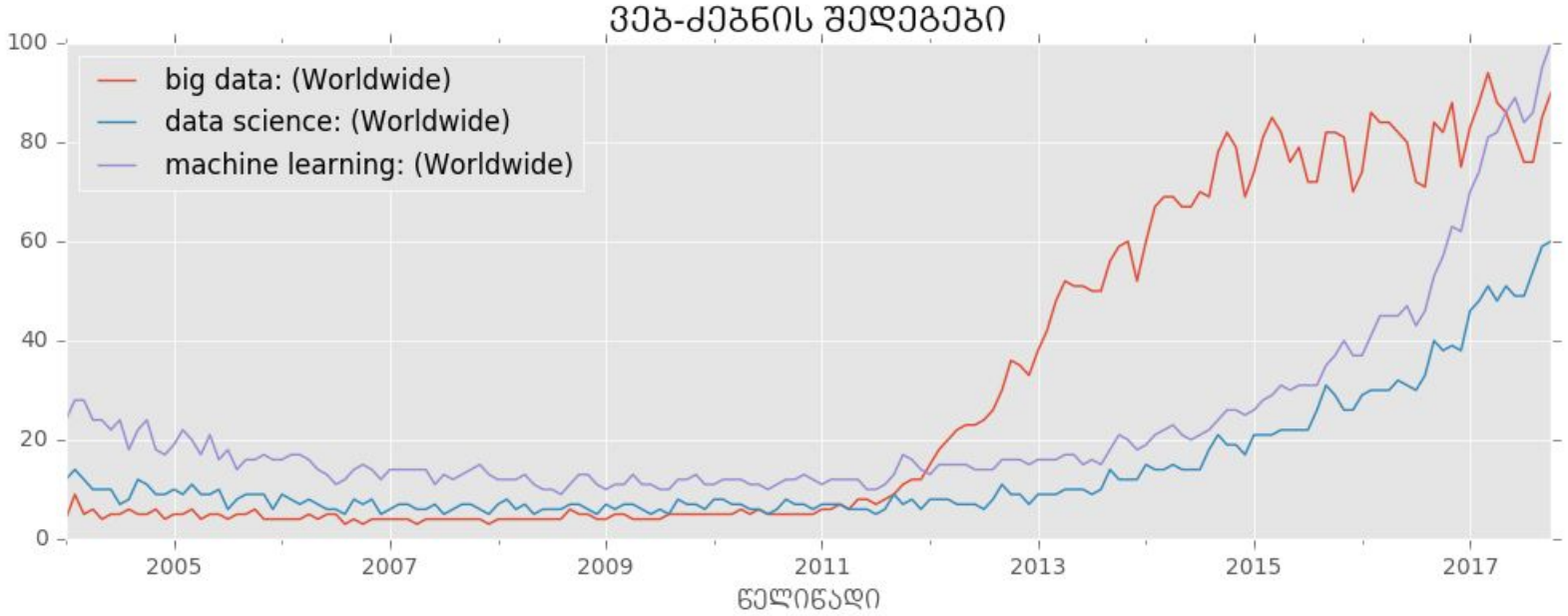
- მონაცემების ავტომატური დამუშავებისა და სასარგებლო ცოდნის მიღების მეთოდები [1]
- დარგი, რომელიც ჯერ კიდევ ჩამოყალიბების პროცესშია [2]

...

შესავალი

მონაცემთა მეცნიერება არის:

- პასუხი მანქანური სწავლებისა და “დიდი მონაცემების” გამოწვევებზე [3]*



* გრაფიკის მონაცემები ეფუძნება ინტერაქტიულ საძიებოს [3] და შესაძლოა დროთა განმავლობაში შეიცვალოს

შესავალი

მონაცემთა მეცნიერების კომპეტენციის სფერო:

- მონაცემთა შეგროვება, გარდაქმნა და გაწმენდა
- მონაცემთა ვიზუალიზაცია
- სტატისტიკური ანალიზი
- მანქანური სწავლება და სხვ.

მონაცემთა მეცნიერების ამოცანები:

- პროგნოზირება
- საიმედოობის შეფასება
- ანომალიათა აღმოჩენა და სხვ.

შესავალი

მონაცემთა მეცნიერს ასევე სჭირდება:

- ფაილებთან მუშაობა
- მონაცემთა ბაზებთან მუშაობა
- სერვისულ ინტერფეისებთან მუშაობა
- განაწილებულ სისტემებთან მუშაობა

შესავალი

მონაცემთა მეცნიერება გამოსადეგია როდესაც:

- შიდა მონაცემები არაა ცენტრალიზებული
  - გროვდება რამდენიმე მონაცემთა ბაზაში და ფაილებში
  - გამოიყენება გარე სერვისები
- მონაცემების მოცულობა და დაგროვების სიჩქარე იზრდება
  - ორგანიზაცია დგას ე.წ. “დიდი მონაცემების” (Big Data) გამოწვევების წინაშე [4]

მონაცემთა პირველადი გამოკვლევა

- რაოდენობრივი შემოწმება
- ნორმალიზაცია
- ვიზუალიზაცია
- მარტივი სტატისტიკური ანალიზი

მონაცემთა პირველადი გამოკვლევა

- რაოდენობრივი შემოწმება
  - სხვადასხვა წყაროების შედარება
  - კატეგორიული ველისაგან სიმრავლის შედგენა
  - კატეგორიების რაოდენობრივი აგრეგაციები
  - რიცხვითი მაჩვენებლების ჯამი და დიფერენციალი
  - ცარიელი ან გამოტოვებული სიდიდეები

მონაცემთა პირველადი გამოკვლევა

- ნორმალიზაცია
  - კატეგორიების დაყვანა ნორმალურ ფორმაზე

მაგ:

სახელწოდება	ენა	...	ფასი
Data Science	ინგლისური	...	25.00
Big Data	ინგლ.	...	29.90
Machine Learning	ინგლისური	...	42.50
Handbuch Statistik	გერ.	...	30.00
Einführung in die Mathematik	გერმანული	...	35.30

მონაცემთა პირველადი გამოკვლევა

- ნორმალიზაცია
  - რიცხვითი სიდიდეების დაყვანა ერთ გარკვეულ ერთეულზე:

მაგ:

სახელწოდება	წონა	...	ფასი
Data Science	1.2	...	25.00
Big Data	1.5	...	29.90
Machine Learning	1.4	...	42.50
Handbuch Statistik	640	...	30.00
Einführung in die Mathematik	900	...	35.30

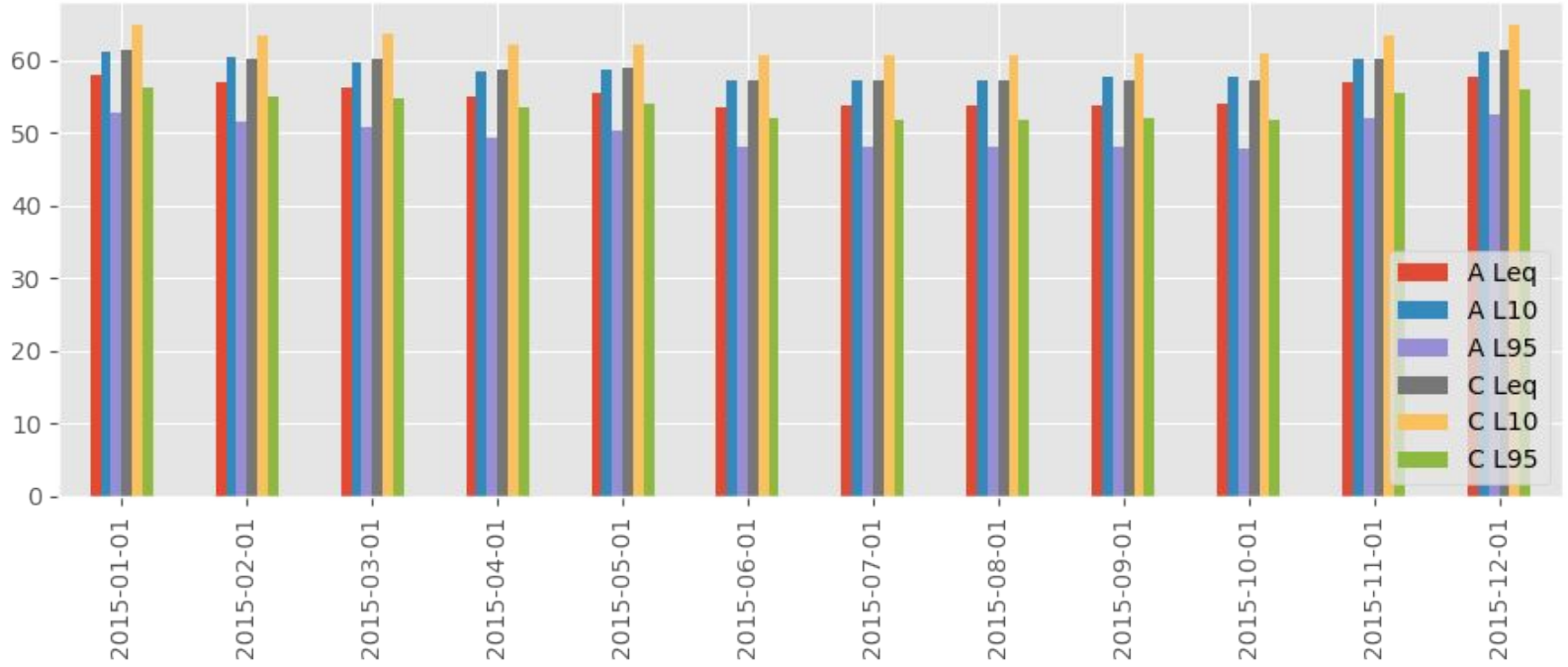
(1 გირვანქა ≈ 453,592 გრ.)

მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - სვეტები
  - წირები
  - წერტილოვანი გრაფიკი
  - სითბური გრაფიკი
  - და სხვ.

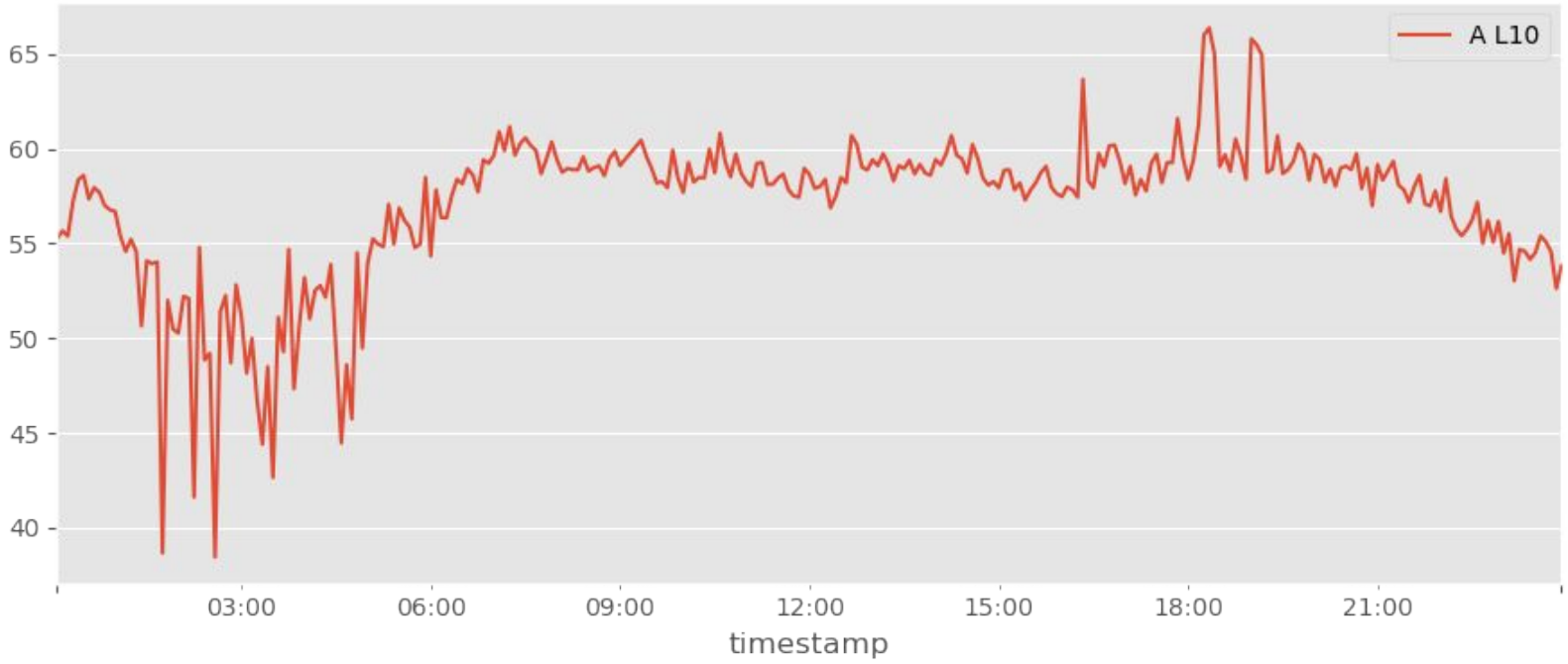
მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - სვეტები - რაოდენობა ან საშუალო მაჩვენებელი



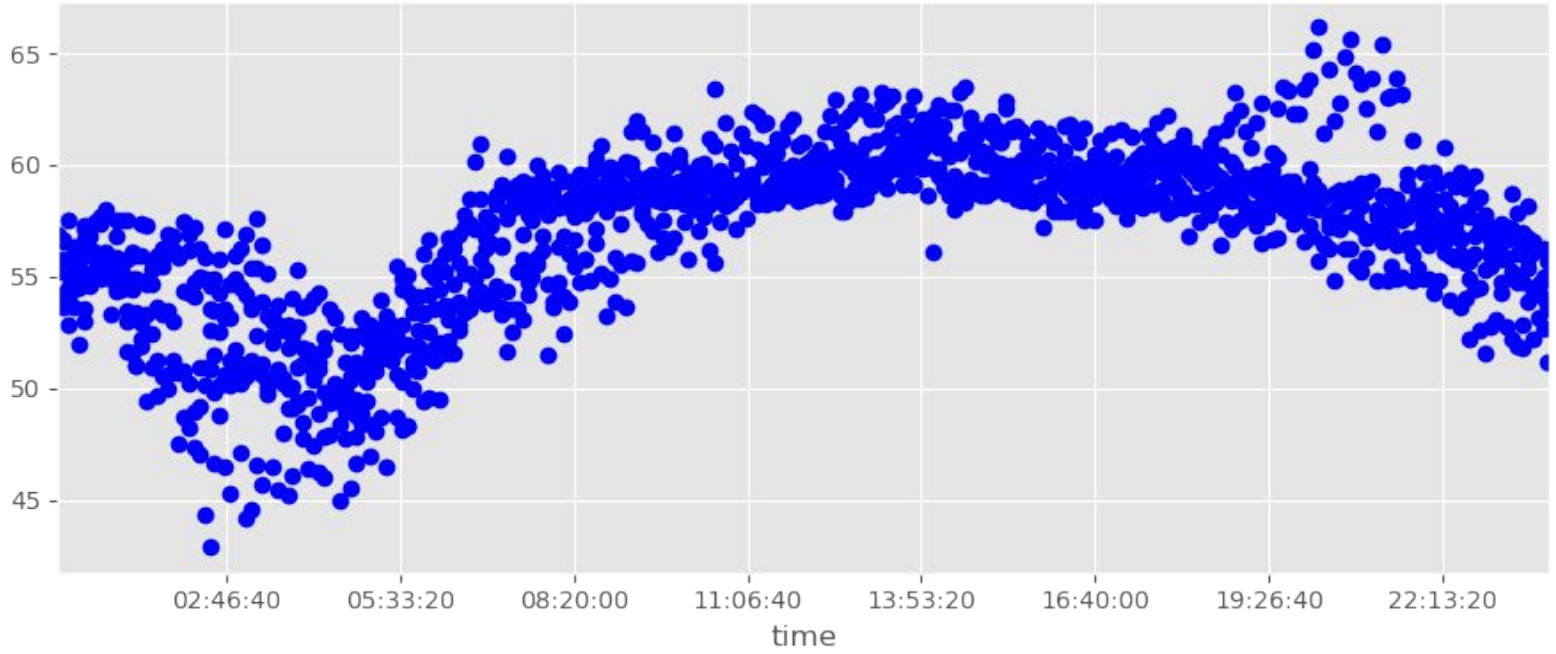
მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - წირები - დროითი მწკრივები



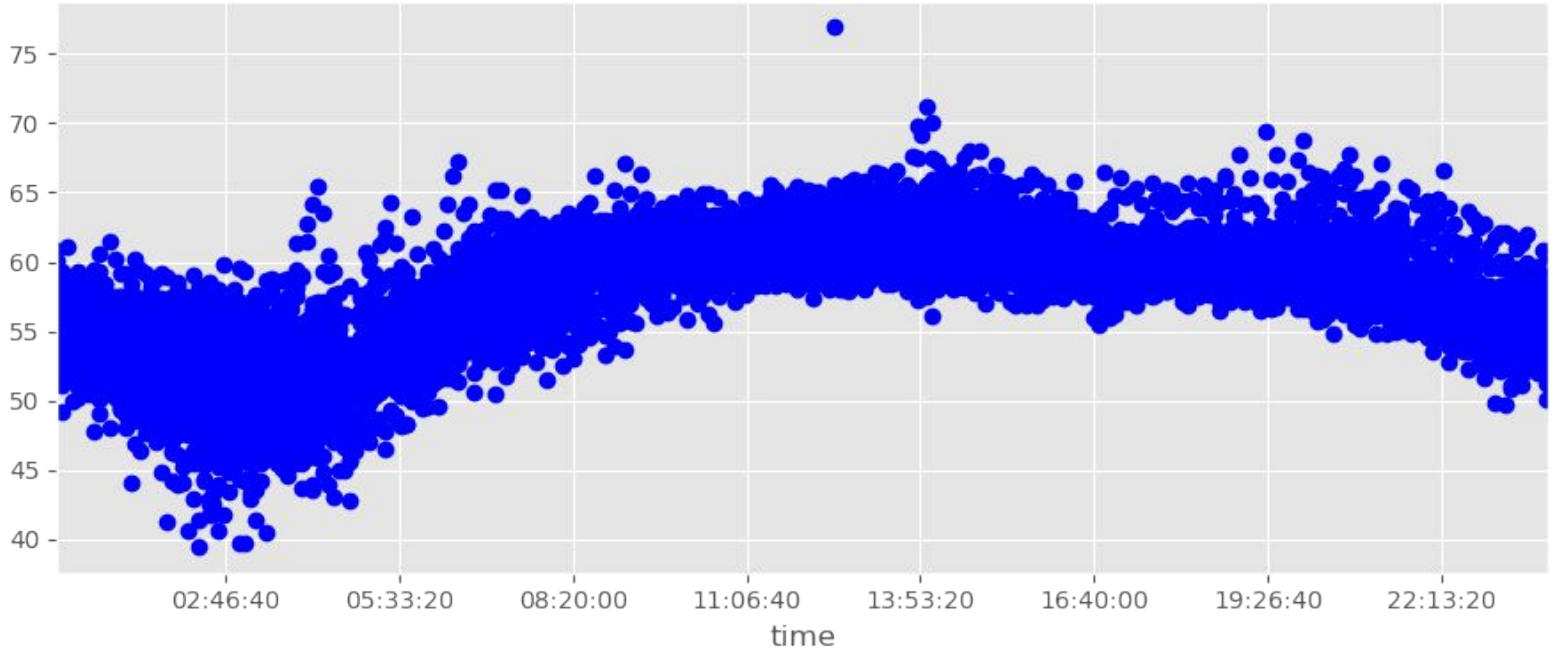
მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - წერტილოვანი გრაფიკი - დროითი მწკრივები, განაწილებები, სიმკვრივე



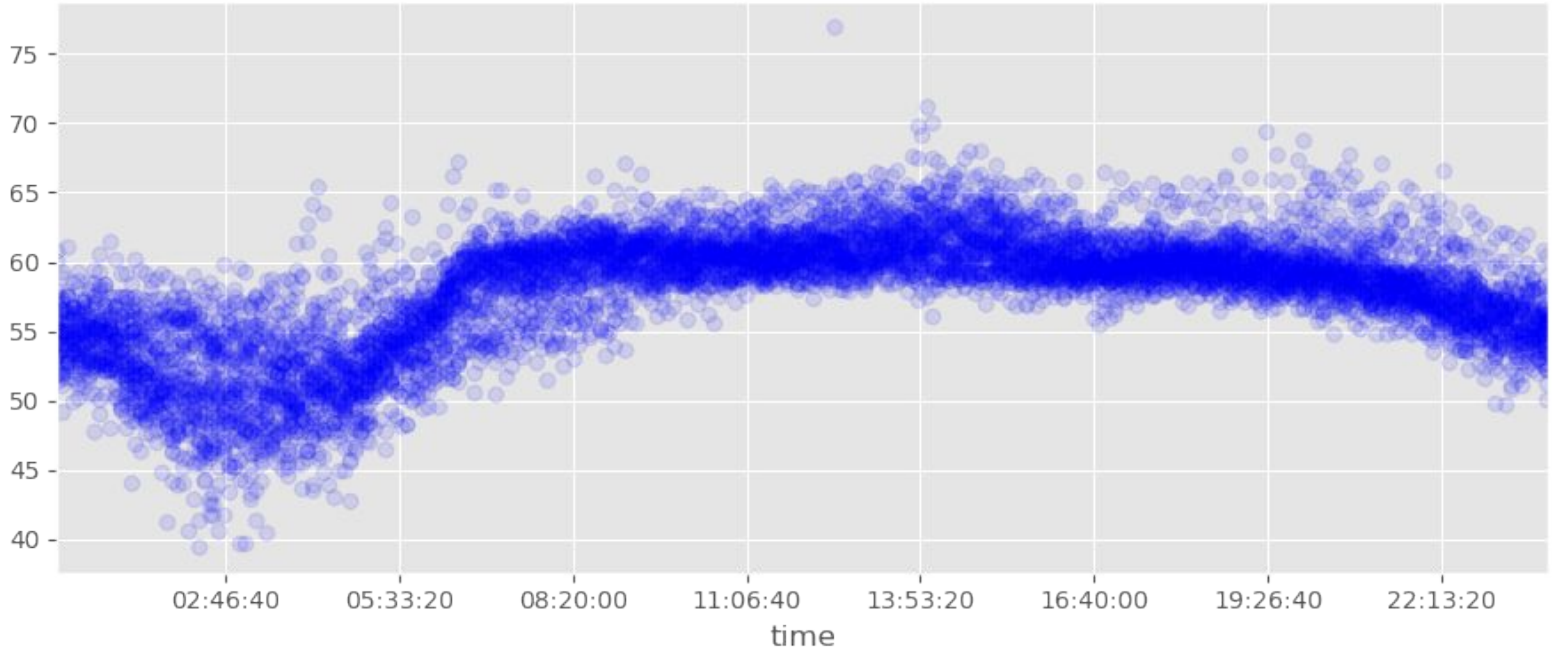
მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - წერტილოვანი გრაფიკი - დროითი მწკრივები, განაწილებები, სიმკვრივე



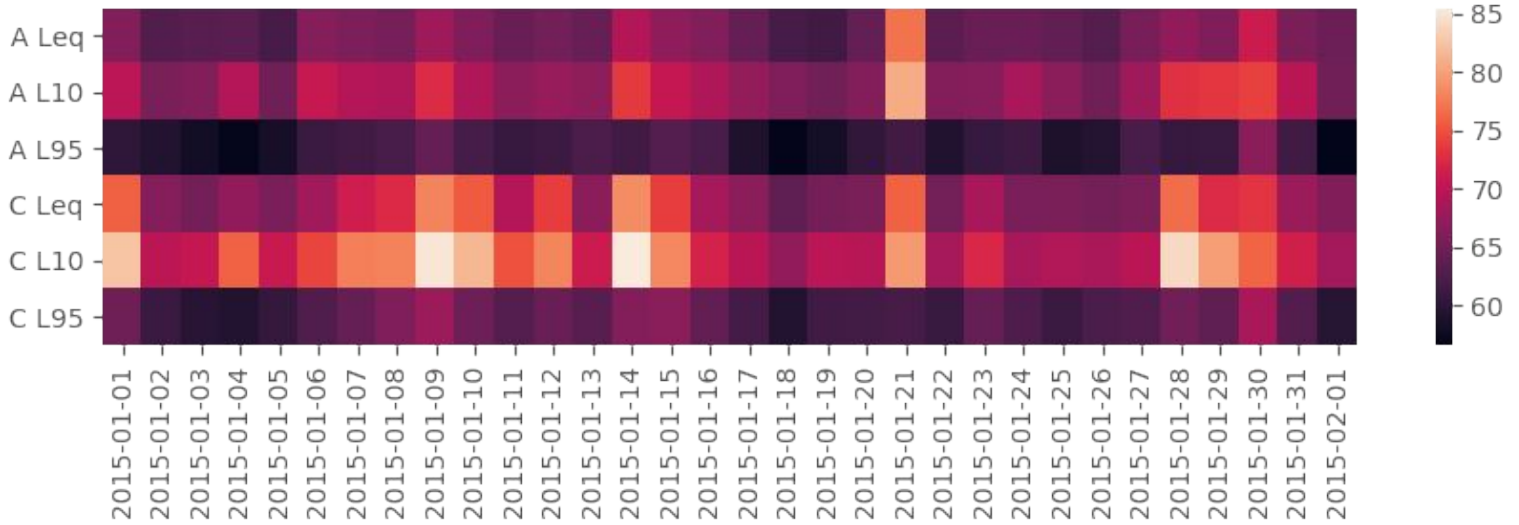
მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - წერტილოვანი გრაფიკი - დროითი მწკრივები, განაწილებები, სიმკვრივე



მონაცემთა პირველადი გამოკვლევა

- ვიზუალიზაცია
  - სითბური გრაფიკი - განაწილებები, სიმკვრივე, ინტენსივობა



მონაცემთა პირველადი გამოკვლევა

- მარტივი სტატისტიკური ანალიზი
  - მინიმუმი და მაქსიმუმი
  - საშუალო არითმეტიკული, სტანდარტული გადახრა
  - მეოთხედები (Quartiles) და მეასედები (Percentiles)
  - დროითი მწკრივის დაჯგუფება საათებად, დღეებად და ა.შ.
    - და მისი ანალიზი ზემოხსენებული მეთოდებით
  - დროითი მწკრივის I და II რიგის დისკრეტ. დიფერენციალი
    - და მათი ანალიზი ზემოხსენებული მეთოდებით
  - ანალიზის პროცესისა და შედეგების ვიზუალიზაცია

მონაცემთა გაწმენდა და გამდიდრება

- ცარიელი და დაუშვებელი ჩანაწერები
- თარიღისა და დროის უნიფიკაცია
- თარიღისა და დროის ანალიზი
- მონაცემთა გამდიდრება და თვისებათა ინჟინერია

მონაცემთა გაწმენდა და გამდიდრება

- ცარიელი და დაუშვებელი ჩანაწერები
  - ტექნიკური თვალისაზრისით
  - დარგობრივი ცოდნიდან გამომდინარე

მონაცემთა გაწმენდა და გამდიდრება

- თარიღისა და დროის უნიფიკაცია
  - სხვადასხვა ფორმატის წაკითხვა
  - გარდაქმნა ერთ ფორმატში
  - 1970 წელი ან შორეული მომავლის თარიღები

მონაცემთა გაწმენდა და გამდიდრება

- თარიღისა და დროის ანალიზი
  - დროის ლოკალიზაცია, “ზაფხულის დრო” (DST) და მსოფლიო კოორდინირებული დრო (UTC)
  - ხდომილებათა სიხშირე და დროითი დიფერენციალის შემოწმება
  - გამოტოვებული (დაკარგული) მონაკვეთები
  - დროითი ამორჩევა და ინტერპოლაცია

მონაცემთა გაწმენდა და გამდიდრება

- მონაცემთა გამდიდრება და თვისებათა ინჟინერია
  - საჯარო მონაცემების დამატება
  - დროითი მაჩვენებლების დაშლა და თვისებად წარმოდგენა
  - კატეგორიათა დაშლა ბინარულ თვისებად - ეკუთვნის თუ არა ჩანაწერი კატეგორიას
  - დისკრეტული დიფერენციალი და ჯამი
  - ორი მაჩვენებლის სხვაობა, ჯამი, ნამრავლი, შეფარდება (დროში სინქრონიზებული)

სტატისტიკური ანალიზი

- მათემატიკური მოლოდინი და სტანდარტული გადახრა
- კორელაციები, კორელაციათა მატრიცა
- ავტოკორელაცია
- კორელაცია დაყოვნებით
- სეზონურობა და ტენდენციები (Trend)
- ხდომილებათა სიხშირის მათემატიკური მოლოდინი და განაწილება

“დემო”

საჯარო მონაცემები - დუბლინის ხმაურის მაჩვენებლები უბნების მიხედვით [5]

დამატებითი დარგობრივი ინფორმაციისათვის იხ. [6]

შეჯამება

- მონაცემების შესახებ ყველა დაშვების ვალიდაცია
- დარღვეული დაშვებებისთვის შესწორების სქემის შემუშავება
- მონაცემთა დამუშავების, გაწმენდისა და გამდიდრების გარშემო გადაწყვეტილებების მიღება ბიზნეს ამოცანიდან გამომდინარე

...

შეჯამება

- მონაცემთა ანალიზი და მომზადება არის ინტერაქტიული და იტერაციული პროცესი
- მანქანური სწავლების თითოეული მოდელისათვის მონაცემთა კრებულის მომზადება მოდელის ტექნიკური საჭიროებების გათვალისწინებით
- მონაცემთა კრებულის უქცევადობა (Immutability) და იდენტიფიკაცია საკონტროლო ჯამებით (Checksums)

ბიბლიოგრაფია

1. https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning
2. https://blog.udacity.com/2018/01/4-types-data-science-jobs.html
3. https://trends.google.com/trends/explore?date=all&q=big%20data,data%20science,machine%20learning
4. https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
5. https://data.smartdublin.ie/dataset/ambient-sound-monitoring-network
6. http://www.acoustic-glossary.co.uk/leq.htm

ზოგადი რეკომენდაცია

1. Allen B. Downey. 2011. Think Stats. O'Reilly Media, Inc. (Free PDF)
2. Joel Grus. 2015. Data Science from Scratch. O'Reilly Media, Inc.
3. Wes McKinney. 2017. Python for Data Analysis, 2nd Edition. O'Reilly Media, Inc.

გმადლობთ ყურადღებისათვის!

რევაზ ტატიშვილი

revaz.tatishvili@gmail.com

https://github.com/rtatishvili

https://www.linkedin.com/in/revaz-tatishvili-79177a71/