

Projekt

Projekti saab teha Individuaalselt või kahekeisi (paaristööna).

Valige nimekirjast andmestik (andmestikud asuvad kaustas Datasets – data.vk.edu.ee kaust Pr).

1. Top 270 Computer Science / Programing Books

https://www.kaggle.com/datasets/thomaskonstantin/top-270-rated-computer-scienceprogramming-books?select=prog_book.csv See andmestik sisaldab loetelu 270 informaatika- ja programmeerimisalasest raamatust. prog_book.csv

2. Top Programming Guru <https://www.kaggle.com/datasets/code19/top-programmingguru?select=video.csv>

On loodud fail (video2021.csv) – andmekogum – 2021 youtube video programmeerimise kohta.

3. Google Play Store Apps <https://www.kaggle.com/datasets/lava18/google-play-store-apps> Andmed rakenduste kohta Play Store'is googleplaystore.csv

Looge fail jupyter notebook.

- Töö alguses peab olema andmestiku üksikasjalik kirjeldus, millised veerud ja mida neis hoitakse, mis tüüpi andmed, mitu rida jne.
- Mõelge välja 5 päringut filtreerimise ja andmete rühmitamisega.
- Uurige ka andmete visualiseerimise teeki, nagu matplotlib või seaborn, või bokeh, või plotly... Koostage 3–5 diagrammi.

Links

[Python Seaborn Tutorial For Beginners: Start Visualizing Data](#)

[Introduction to Plotting with Matplotlib in Python](#)

- Looge HTML või PDF fail jupyter notebook'ist. Failis peavad olema skriptid (python code) ja nende kirjeldused (markdown).

Näiteks,

Fail programmeerimiskeelte kohta. Link: <https://www.kaggle.com/datasets/fernando/countries-of-the-world>

Andmekogumi kohta (dataset)

Tavaline küsimus informaatika ja tarkvaraarenduse algajatele ja kursis olevatele inimestele: milline programmeerimiskeel on parim ja/või populaarseim? Ühetähenduslikku vastust on väga raske anda, sest on näiliselt määramatu arv näitajaid, mille järgi saab määrata "parimat" või "kõige populaarsemat" programmeerimiskeelt.

Üks selliseid näitajaid, mida saab kasutada "populaarse" programmeerimiskeele määramiseks, on projektide ja failide arv, mis on loodud seda programmeerimiskeelt kasutades. Kuna GitHub on kõige populaarsem üldkasutatav ühiskasutus- ja failijagamisplatvorm, on GitHubi hoidlate, projektide ja probleemide jaoks kasutatavate keelte analüüs hea keele populaarsuse näitaja.

Andmestik sisaldab statistikat programmeerimiskeelte kohta, mida kasutatakse GitHubi projektide loomisel. Andmed on aastatest 2011-2022.

Allikas: <https://www.kaggle.com/datasets/isaacwen/github-programming-languages-data?select=prs.csv>

Fail on salvestatud kaustas data

```
In [ ]: df = pd.read_csv('data/projects_github.csv')
df.shape
```

```
Out [ ]: (3462, 4)
```

3462 rida ja veergu DataFrame-is

Näide: vt. fail Project_Kaarel_Kilki.html