

Juhend statistikaameti andmebaasi kasutamiseks R kaudu

Käesolev juhend annab ülevaate statistikaameti eestikeelse andmebaasi <https://andmed.stat.ee/et/stat> kasutamisest tarkvaraga R.

Andmebaasi kasutamiseks soovime paketti pxweb. Selle installimiseks `install.packages("pxweb")` ja vajadusel uuendage ka varasemalt installitud pakette.

```
library(pxweb)
```

Nimetatud pakett võimaldab andmetele ligipääsu erinevatel viisidel, millest kolmel peatume lähemalt.

Andmed interaktiivselt

Kasutades käsku `pxweb_interactive()` on võimalik andmebaasis teha valikuid olemasolevate alamjaotuste kaupa. Esmalt palutakse valida kõige laiem teemade jaotus (vt järgnevat ekraanipilti), seejärel täpsustused ning lõpuks tulemuste esitamise viis. Käsus tuleb ette anda andmebaasi aadress

```
pxweb_interactive("https://andmed.stat.ee/api/v1/et/stat")
```

```
=====
R PXWEB: Content of 'andmed.stat.ee'
        at '/api/v1/et/stat'
=====
[ 1 ] : Eri valdkondade statistika
[ 2 ] : Keskkond
[ 3 ] : Majandus
[ 4 ] : Põllumajandusloendus
[ 5 ] : Rahva ja eluruumide loendus
[ 6 ] : Rahvastik
[ 7 ] : Sotsiaalelu
=====
Enter your choice:
('esc' = Quit, 'b' = Back, 'i' = Show id)
1:
```

Andmed konkreetse tabeli kohta, mille numbrit me teame

Juhul, kui on vaja andmeid juba teada oleva numbriga tabelist, võib pöörduda ka otse selle tabeli poole. Järgnevas näites soovime saada kuiseid andmeid erinevate tegevusalade keskmiste palkade kohta tabelist PA006. Selleks tuleb eelnevalt kasutatud aadressi täiendada tabeli numbriga. Samuti tuleb päringus ära näidata, et milliseid näitajaid, tegevusalasid ja milliseid vaatlusperioode (aastaid ja kuid) me oma päringusse võtta tahame. Viimaste valikuks saab kasutada kas json formaadis koodi või Ri enda liste.

Päring json formaadis koodiga

Päringu parameetrite etteandmiseks json formaadis koodiga tuleks soovitud näitajad ja aastad koos soovitud väärtustega eraldi failina salvestada. Näiteks, kui soovime vaatlusperioodiks valida alates 2018, 2019 ja 2020. aasta kuus esimest kuud, tegevusalad “Majutus ja toitlustus”, “Töötlev tööstus” ning “Ehitus” ja keskmise brutokuupalga, siis peaks kood olema selline:

```
{
  "query": [
    {
      "code": "Tegevusala",
      "selection": { "filter": "item",
        "values": [ "C","F","I" ]
      }
    },
    {
      "code": "Näitaja",
      "selection": { "filter": "item",
        "values": [ "D11_EMPL"
      ]
    }
  ],
  {
    "code": "Vaatlusperiood",
    "selection": { "filter": "item",
      "values": [
        "2018M1",
        "2018M2",
        "2018M3",
        "2018M4",
        "2018M5",
        "2018M6",
        "2019M1",
        "2019M2",
        "2019M3",
        "2019M4",
        "2019M5",
        "2019M6",
        "2020M1",
        "2020M2",
        "2020M3",
        "2020M4",
        "2020M5",
        "2020M6" ]
    }
  }
],
  "response": {
    "format": "json"
  }
}
```

Järgnevas näites on see fail salvestatud kodukataloogi nimega naide.json. Tulemus antakse Ri listina.

```
px_andmed_json <- pxweb_get(url = "https://andmed.stat.ee/api/v1/et/stat/PA006",  
                           query = "naide.json")
```

Andmete saamiseks Ris data.frame formaati saab kasutada järgmist käsku. Kõigepealt antakse ette eelmise käsuga saadud objekt. Järgmised käsuosad täpsustavad, kas kasutada tabelis väärtuste silte tekstina (praegu "text") või koodidena (saaks ka valida tüübiks "code").

```
px_df_json <- as.data.frame(px_andmed_json, column.name.type = "text", variable.value.type = "text")
```

Päring Ri listi formaadis koodiga

Päringu query osas saaks tingimusi ette anda ka Ri listina. Eraldi peavad olema kajastatud kõik tabelis esinevad alamjaotused. List 2020. aasta esimeste kuude kohta näeks välja selline

```
px_query_list1 <- list("Tegevusala" = c("C", "F", "I"),  
                      "Näitaja" = "D11_EMPL",  
                      "Vaatlusperiood" = c("2020M1", "2020M2", "2020M3", "2020M4", "2020M5", "2020M6"))
```

Alternatiivina oleks võimalik koostada päring selliselt, et võetakse kõik olemasolevad aastad ja näitajad. Sel juhul tuleb päringusse märkida tärn. Kõik tabelis esinevad alamjaotused (Tegevusala, Näitaja ja Vaatlusperiood) peavad aga endiselt olema olema.

```
px_query_list2 <- list("Tegevusala" = c("C", "F", "I"),  
                      "Näitaja" = c("*"),  
                      "Vaatlusperiood" = c("*"))
```

Päringus endas tuleb siis ette anda eelnevalt moodustatud listi objekt.

```
px_andmed_list1 <- pxweb_get(url = "https://andmed.stat.ee/api/v1/et/stat/PA006",  
                           query = px_query_list1)
```

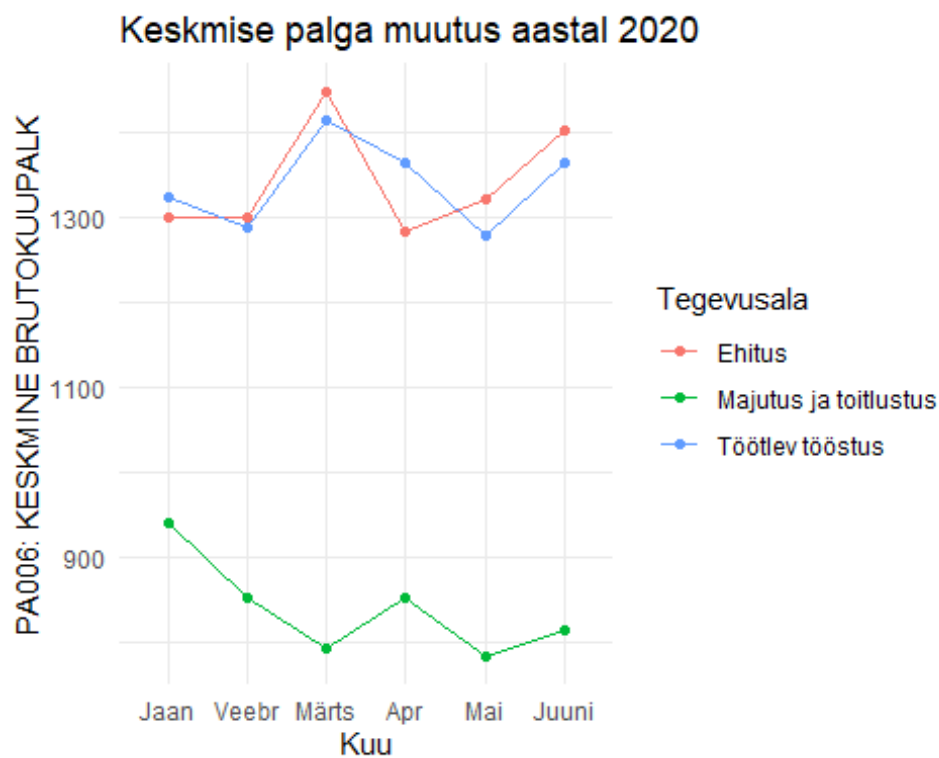
Andmetabeli saaks listiga tehtud päringu tulemustest kokku panna sarnase käsuga kui eelnevaltki

```
px_df_list1 <- as.data.frame(px_andmed_list1, column.name.type = "text", variable.value.type = "text")
```

Saadud andmed on pikkformaadis, mis võimaldab sobivaid aastaid ja näitajaid analüüsiks kergesti välja filtreerida ja kasutada näiteks joonistel.

```
library(tidyverse)

px_df_list1 %>%
  mutate(Kuu = fct_relevel(substring(Vaatlusperiood, 6), "Jaan", "Veebr", "Märts", "Apr", "Mai", "Juuni")) %>%
  ggplot(aes(x = Kuu, y = `PA006: KESKMINE BRUTOKUUPALK`, group = Tegevusala)) +
  geom_line(aes(color = Tegevusala)) +
  geom_point(aes(color = Tegevusala)) +
  ggtitle("Keskmise palga muutus aastal 2020") +
  theme_minimal()
```



Metaandmed

Päringuid lihtsustab ilmselt see, et andmebaasist on võimalik vaadata ka tabelite ja koodide kohta käivaid metaandmeid.

Andmebaasi alajaotusi saab vaadata näiteks käskudega

```
pxweb_get("https://andmed.stat.ee/api/v1/et/stat")

## PXWEB LEVELS
## eri-valdkondade-statistika (1): Eri valdkondade statistika
## keskkond (1): Keskkond
## majandus (1): Majandus
## pm-loendus (1): Põllumajandusloendus
## rahvaloendus (1): Rahva ja eluruumide loendus
## rahvastik (1): Rahvastik
## sotsiaalelu (1): Sotsiaalelu

pxweb_get("https://andmed.stat.ee/api/v1/et/stat/majandus")

## PXWEB LEVELS
## ehitus (1): Ehitus
## energieetika (1): Energeetika
## ettevõtete-majandusnäitajad (1): Ettevõtete majandusnäitajad
## hinnad (1): Hinnad
## infotehnoloogia (1): Infotehnoloogia ja side
## jahindus (1): Jahindus
## kalandus (1): Kalandus
## kinnisvara (1): Kinnisvara
## majandusüksused (1): Majandusüksused
## metsamajandus (1): Metsamajandus
## palk-ja-toojeukulu (1): Palk ja tööjõukulu
## pellumajandus (1): Põllumajandus
## rahandus (1): Rahandus
## rahvamajanduse-arvepidamine (1): Rahvamajanduse arvepidamine
## sisekaubandus (1): Sisekaubandus
## teadus-tehnoloogia-innovatsioon (1): Teadus. Tehnoloogia. Innovatsioon
## teenindus (1): Teenindus
## transport (1): Transport
## turism-ja-majutus (1): Turism, majutus ja toitlustus
## toostus (1): Tööstus
## valiskaubandus (1): Väliskaubandus
```

Vastaval tasemel näidatud alajaotuse nime saab kasutada järgmise taseme päringu tegemisel.

Konkreetse tabeli metaandmete vaatamiseks saab kasutada käsku

```
px_meta <- pxweb_get("https://andmed.stat.ee/api/v1/et/stat/PA006")
```

Tekkinud objekt on jälle esitatud listina.

```
px_meta
```

```
## PXWEB METADATA
## PA006: KESKMINE BRUTOKUUPALK TEGEVUSALARÜHMA JÄRGI (KUUD)
## variables:
## [[1]] Tegevusala: Tegevusala
## [[2]] Näitaja: Näitaja
## [[3]] Vaatlusperiood: Vaatlusperiood
```

Vaatamaks tunnuseid, mida meie tabel sisaldab võib uurida osa variables

```
px_meta$variables
```

Siit selgub ka, et tegevusalad on esitatud esimeses ja näitajad teises elemendis. Sedasama teist elementi täpsemalt uurides saab ka näha Näitajate koode võrdlevalt koos tekstidega.

```
px_meta$variables[[2]]
```

```
## $code
## [1] "Näitaja"
##
## $text
## [1] "Näitaja"
##
## $values
## [1] "D11_EMPL"      "D11_EMPL_CJAN" "D11_EMPL_CMTH"
##
## $valueTexts
## [1] "Keskmine brutokuupalk"
## [2] "Keskmise brutokuupalga muutus võrreldes jaanuariga"
## [3] "Keskmise brutokuupalga muutus võrreldes eelmise kuuga"
##
## $elimination
## [1] FALSE
##
## $time
## [1] FALSE
```

Head kasutamist kõigile!