Veebi kraapimine

Risto Hinno

Friday, June 26, 2015

Andmete kraapimine veebist

Sissejuhatus

Eva "Usin" Masin on esimeses praktikumis kohatud Mati "Raha" Masina vastand. Talle meeldib rutiinsus, andmete tuim kopeerimine ja sisestamine. Vabal ajal meeldib talle lugeda romaanisarja "Tõde ja õigus" - eelmine nädal luges ta kokku tähekombinatsiooni "pa" esinemissageduse. Homme pärast tööd jätkab ta "pb" esinemissageduse leidmisega.

Kahjuks oli eelnev tekst fiktsioon ning eva-laadsed kopeerijad surid välja koos neandertaallastega. Selles praktikumis vaatame, kuidas R-is ellu äratada tehis-Eva, kes oskab veebilehtedelt automaatselt infot eraldada ja selle transformeerida struktureeritud andmestikuks.

Täpsemalt, uurime

- kuidas eraldada Riigikogu hääletamistulemusi,
- kuidas eraldada Postimehe uudiste pealkirju,
- kuidas eraldada ilmajaama vaatlusandmeid.

Kaks esimest ülesannet õpetavad paketi rvest funktsionaalsust ja annavad sissejuhatuse veebikraapimisse minimalistlike veebilehtede põhjal.

Ülesanne 1 (2 punkti) - CSS id

- Eralda html koodis sinisena olev tekst muutujasse tekst. Kasuta paketti rvest.
- Vastava html koodiga saad mängida siin.
- Loe lühiülevaadet, millest koosnevad veebilehed.
- Uuri paketi rvest minimalistlikku näidet.
- Minimalistliku näite põhjal peaksid oskama eraldada lähtekoodis olevad 4 lõiku. Et eraldada lõik, kus id="p01", pead teadma, kuidas CSS-is tähistatakse id-sid. Suur vihje on olemas eelneva html koodi < style > blokis. Abiks võib-olla ka CSS selektorite interaktiivne testnäide.

```
library(rvest)
html_source ="http://andmeteadus.github.io/examples/html1.html"
page = html(html_source)
tekst=page %>%
    html_node("p#p01") %>%
    html_text()
tekst
```

```
## [1] "I am different."
```

Ülesanne 2 (2 punkti) - CSS class

Eralda html koodis punaselt olev tekst muutujasse tekst. Kasuta paketti rvest.

Vastava html koodiga saad mängida siin.

Lõpptulemuse peaks olema selline: tekst = c("I am different.", "I am different too.")

```
html_source ="http://andmeteadus.github.io/examples/html2.html"
page = html(html_source)
tekst2=page %>%
    html_nodes("p.error") %>%
    html_text()
tekst2
```

```
## [1] "I am different."     "I am different too."
```

Ülesanne 3 (2 punkti)

Eralda Riigikogu hääletamistulemuste veebilehe html lähtekoodist, mitu saadikut hääletas kooseluseaduse eelnõu:

- poolt
- vastu
- oli erapooletu
- ei hääletanud

Praktikumis tutvusime, kuidas brauseri veebiarendus tööriistadega leida üles lähtekoodist vajalikud kohad. Variandid olid:

- Chrome'is vajuta parem klikk ja "inspekteeri elementi". Alernatiivid on klahvikombinatsioon Ctrl + Shift + I või klahv F12. vahendiga selectorgadget
- Need muudavad lähtekoodis õige klassi, id või sildi leidmise oluliselt lihtsamaks. Mõnes olukorras on kasulikum üks variant, mõnes teine.

```
html_source ="http://www.riigikogu.ee/tegevus/tooulevaade/haaletused/haaletustulemused-kohalolekukontro
page = html(html_source)
haaletus=page %>%
    html_nodes("li a span") %>%
    html_text()
haaletus
```

```
## [1] "40"  "38"  "0"   "10"  " 88" " 13"
```

Ülesanne 4 (2 punkti)

Eralda kooseluseaduse eelnõu hääletamistulemuste veebilehe html lähtekoodist andmetabel, kus on 101 rida ning tunnused nr, nimi, otsus, fraktsioon.

Vihje: kasuta funktsiooni html_table

Kirjuta vastav kood funktsiooniks extract_table (seda funktsiooni läheb vaja järgmises ülesandes, kus eraldad kõigi Riigikogu XII hääletuste kohta vastava tabeli). Sisendiks on kas veebilehe url, faili lokaalne asukoht või sõne. Funktsioon peab tagastama vastava data.frame-i (pane tähele, et su funktsioon ei tagastaks listi, milles on üks data.frame).

```
library(knitr)
html_source ="http://www.riigikogu.ee/tegevus/tooulevaade/haaletused/haaletustulemused-kohalolekukontro
page = html(html_source)

haaletus=page %>%
    html_nodes(".table.table.table-striped.full-bars") %>%
    html_table()
#siit saan data frame 101 liikme, erkonna ja hööletustulemusega, nr-d võtan
#zipitud tabelitest, kuna vahepeal on kodulehte muudetud
sub=haaletus[1]
sub2=as.data.frame(sub)
#puhastan otsuse poolt ja asendan tabelisse
junn=gsub( " ", "",sub2$Otsus)
junn2=gsub( "\n\n", "",junn)
junn3=gsub( "^Poolt", "",junn2)
junn3=gsub( "^Vastu", "",junn3)
junn3=gsub( "^EiHääletanud", "",junn3)
junn3=gsub( "^Puudub", "",junn3)
junn3=gsub( "EiHääletanud", "Ei Hääletanud",junn3)
sub2$Otsus=junn3
kable(sub2)
```

Nimi	Otsus	Fraktsioon
Arto Aas	Poolt	Eesti Reformierakonna fraktsioon
Jaak Aaviksoo	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Rein Aidma	Poolt	Eesti Reformierakonna fraktsioon
Annely Akkermann	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Jaak Allik	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Peep Aru	Vastu	Eesti Reformierakonna fraktsioon
Maimu Berg	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Deniss Boroditš	Vastu	Fraktsiooni mittekuuluvad Riigikogu liikmed
Enn Eesmaa	Puudub	Eesti Keskerakonna fraktsioon
Eldar Efendijev	Vastu	Eesti Keskerakonna fraktsioon
Ene Ergma	Puudub	Isamaa ja Res Publica Liidu fraktsioon
Igor Gräzin	Vastu	Eesti Reformierakonna fraktsioon
Margus Hanson	Puudub	Eesti Reformierakonna fraktsioon

Nimi	Otsus	Fraktsioon
Aare Heinvee	Ei Hääletanud	Eesti Reformierakonna fraktsioon
Andres Herkel	Vastu	Fraktsiooni mittekuuluvad Riigikogu liikmed
Remo Holsmer	Poolt	Eesti Reformierakonna fraktsioon
Kaia Iva	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Jüri Jaanson	Poolt	Eesti Reformierakonna fraktsioon
Tatjana Jaanson	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Kalle Jents	Poolt	Eesti Reformierakonna fraktsioon
Tõnu Juul	Vastu	Eesti Reformierakonna fraktsioon
Etti Kagarov	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Lembit Kaljuvee	Poolt	Fraktsiooni mittekuuluvad Riigikogu liikmed
Kalev Kallemets	Poolt	Eesti Reformierakonna fraktsioon
Kalev Kallo	Vastu	Eesti Keskerakonna fraktsioon
Siim Kiisler	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Aivar Kokk	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Mihhail Korb	Vastu	Eesti Keskerakonna fraktsioon
Valeri Korb	Vastu	Eesti Keskerakonna fraktsioon
Siret Kotka	Puudub	Eesti Keskerakonna fraktsioon
Kalev Kotkas	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Tõnis Kõiv	Poolt	Eesti Reformierakonna fraktsioon
Kalvi Kõva	Ei Hääletanud	Sotsiaaldemokraatliku Erakonna fraktsioon
Kalle Laanet	Vastu	Fraktsiooni mittekuuluvad Riigikogu liikmed
Lauri Laasi	Poolt	Eesti Keskerakonna fraktsioon
Rein Lang	Puudub	Eesti Reformierakonna fraktsioon
Peeter Laurson	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Tarmo Leinatamm	Puudub	Eesti Reformierakonna fraktsioon
Heimar Lenk	Puudub	Eesti Keskerakonna fraktsioon
Kalev Lillo	Puudub	Eesti Reformierakonna fraktsioon
Väino Linde	Vastu	Eesti Reformierakonna fraktsioon
Tiina Lokk-Tramberg	Poolt	Eesti Reformierakonna fraktsioon
Inara Luigas	Puudub	Fraktsiooni mittekuuluvad Riigikogu liikmed
Lauri Luik	Poolt	Eesti Reformierakonna fraktsioon
Rait Maruste	Poolt	Eesti Reformierakonna fraktsioon
Mart Meri	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Kristen Michal	Poolt	Eesti Reformierakonna fraktsioon
Marko Mihkelson	Ei Hääletanud	Isamaa ja Res Publica Liidu fraktsioon
Marianne Mikko	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon

Nimi	Otsus	Fraktsioon
Jüri Morozov	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Meelis Mälberg	Poolt	Eesti Reformierakonna fraktsioon
Tarmo Mänd	Vastu	Eesti Reformierakonna fraktsioon
Eiki Nestor	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Erki Nool	Puudub	Isamaa ja Res Publica Liidu fraktsioon
Liisa-Ly Pakosta	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Kalle Palling	Poolt	Eesti Reformierakonna fraktsioon
Tõnis Palts	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Juhan Parts	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Heljo Pikhof	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Barbi Pilvre	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Marko Pomerants	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Mati Raidma	Poolt	Eesti Reformierakonna fraktsioon
Laine Randjärv	Poolt	Eesti Reformierakonna fraktsioon
Valdo Randpere	Poolt	Eesti Reformierakonna fraktsioon
Rein Randver	Puudub	Sotsiaaldemokraatliku Erakonna fraktsioon
Jüri Ratas	Ei Hääletanud	Eesti Keskerakonna fraktsioon
Urmas Reinsalu	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Mailis Reps	Puudub	Eesti Keskerakonna fraktsioon
Aivar Riisalu	Vastu	Fraktsiooni mittekuuluvad Riigikogu liikmed
Reet Roos	Puudub	Isamaa ja Res Publica Liidu fraktsioon
Aivar Rosenberg	Ei Hääletanud	Eesti Reformierakonna fraktsioon
Paul-Eerik Rummo	Poolt	Eesti Reformierakonna fraktsioon
Karel Rüütli	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Indrek Saar	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Andrus Saare	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Helir-Valdor Seeder	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Andre Sepp	Poolt	Eesti Reformierakonna fraktsioon
Sven Sester	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Priit Sibul	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Kadri Simson	Poolt	Eesti Keskerakonna fraktsioon
Imre Sooäär	Poolt	Eesti Reformierakonna fraktsioon
Mihhail Stalnuhhin	Vastu	Eesti Keskerakonna fraktsioon
Neeme Suur	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Aivar Sõerd	Vastu	Eesti Reformierakonna fraktsioon
Olga Sõtnik	Poolt	Eesti Keskerakonna fraktsioon

Nimi	Otsus	Fraktsioon
Jaanus Tamkivi	Ei Hääletanud	Eesti Reformierakonna fraktsioon
Tarmo Tamm	Vastu	Eesti Keskerakonna fraktsioon
Tiit Tammsaar	Ei Hääletanud	Sotsiaaldemokraatliku Erakonna fraktsioon
Priit Toobal	Poolt	Eesti Keskerakonna fraktsioon
Terje Trei	Ei Hääletanud	Eesti Reformierakonna fraktsioon
Margus Tsahkna	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Marika Tuus-Laul	Ei Hääletanud	Eesti Keskerakonna fraktsioon
Toomas Tõniste	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Urbo Vaarmann	Vastu	Eesti Keskerakonna fraktsioon
Ken-Marti Vaher	Vastu	Isamaa ja Res Publica Liidu fraktsioon
Rainer Vakra	Poolt	Fraktsiooni mittekuuluvad Riigikogu liikmed
Rannar Vassiljev	Poolt	Sotsiaaldemokraatliku Erakonna fraktsioon
Viktor Vassiljev	Vastu	Eesti Keskerakonna fraktsioon
Vladimir Velman	Vastu	Eesti Keskerakonna fraktsioon
Peeter Võsa	Vastu	Eesti Keskerakonna fraktsioon
Jaan Õunapuu	Ei Hääletanud	Sotsiaaldemokraatliku Erakonna fraktsioon

Funktsiooni ma ei tee, kuna Riigikogu kodulehte on uuendatud ja funktsioon ei töötaks järgnevate asjade peal.

Ülesanne 5 (5 punkti) - andmestiku ehitamine

Ülesandes 4 tegid läbi Riigikogu saadikute hääletamistulemuste eraldamise kooseluseaduse korral. Failis htmls.zip on olemas veebilehed kõigi Riigikogu XII hääletuste kohta. Sinu ülesandeks on koostada andmetabel, kus ridades on Riigikogu saadiku nimi ja veergudes kõik hääletamiskorrad. Seda andmestikku läheb vaja järgmises praktikumis, kus uurime hääletamismustreid.

Kõigepealt paki lahti zip fail ning loe R-i sisse kõigi html failide nimed.

Näpunäide: Järgnev kood loeb sisse kõik muutujas filenames olevad csv andmestikud ning tekitab neist listi.

list_of_dataframes = list() for(i in 1:length(filenames)){ temp = read.csv(filenames[i]) list_of_dataframes[[i]] = temp }

Praegu pole sul read.csv käsuga midagi peale hakata, sest tegeleme html failidega. Kasuta ülesandes 4 kirjutatud funktsiooni extract_table. Eelneva for-tsükli asemel võid kasutada funktsiooni lapply.

Lisa igal tsükli sammul andmestikule hääletuse indeks või muu identifikaator. Näiteks temp$haaletus = i.

Nüüdseks peaksid olema saanud listi, mille elementideks on erinevad andmetabelid (kõiki faile kasutades peaks nende koguarv olema 1845). Tee nendest andmetabelitest üks suur (pikk) andmetabel, paigutades need üksteise otsa. Seda aitab teha paketi dplyr funktsioon rbind_all. Tulemuseks peaksid saama andmetabeli, mille ridade arv on 101 * "sinu kasutatud failide arv".

Muuda pikk andmetabel laiaks. Seda aitab teha paketi reshape2 käsk dcast. Uuri funktsiooni dcast minimalistlikku näidet siit.

Kui kõik eelnev töötab, tee eelnev läbi kõikide html failidega. Ära kohku, kui kõikide html tabelite eraldamisega läheb aega 5 minutit või rohkem.

Soovitus: Kui oled eelneva ühe korra läbi teinud, pole vaja knitri raporti genereerimisel enam sedasama korrata. Saadud andmetabeli saad endale salvestada käsuga save(andmed, file="riigikogu.RData"). Raportis võid muuta vastava koodiploki eval=FALSE.

```
#funktsioon kõige pealt
extract_table=function(url) {
    html_source =url
    page = html(html_source)
    tabel=page %>%
        html_nodes("table.List") %>%
        html_table()%>%
        as.data.frame()
    tabel
}

#loen kõik failid sisse
filenames =list.files("./data/data", pattern = "*.html", full.names=TRUE)

list_of_dataframes = list()
for(i in 1:length(filenames)){
    temp = extract_table(filenames[i])
    temp$haaletus = i
    list_of_dataframes[[i]] = temp
}
#keevitan kokku
library(dplyr)
andmed=rbind_all(list_of_dataframes)

#Muudan pika andmetabeli laiaks
library(reshape2)
andmed_lai=dcast(andmed, Nimi~haaletus, value.var = "Otsus")
```

Ülesanne 6

Tagasta kõik Postimehe esilehe uudiste pealkirjad (joonisel näidatud kollasega).

Ära kurvasta, kui sa ei saa absoluutselt kõiki pealkirju, 97% on praegu piisav.

```
url="http://www.postimees.ee/"
page = html(url)
tekst=page %>%
    html_nodes(".frontHeading") %>%
    html_text()
#puhastame
tekst_puhas=gsub("\n", "", tekst)
#palkirjade lõppu korjab ka kommentaaride arvu, puhastame need välja
tekst_puhas=gsub("\\d*$", "", tekst_puhas)
#eemaldame tühjad stringid
tekst_puhas=tekst_puhas[tekst_puhas != ""]
kable(tekst_puhas)
```

Ateena ei nõustunud kreeditoride ettepanekuga

Turulettidele jõudsid Eesti maasikad

Suri egüptoloog Sergei Stadnikov

Iisrael kutsub Prantsuse juute riigist Iisraeli põgenema

USA ülemkohus seadustas homoabielud kogu riigis

Võõrsil mänginud Kalju kaotas Viljandile

Tartus töötab suvel taas ranna- raamatukogu

Lõbus video: kontorirott teeskleb, et ta valiti NBAsse ja petab Manhattanil kõik ära

Iisrael väljendas Vatikani-Palestiina leppe üle «kahetsust»

Kopenhaagenisse püstitatakse uus hiiglaslik mošee

Islamiriigi rünnakus Kobane'le on hukkunud juba 146 tsiviilisikut

Galerii: rünnak Kuveidi mošeele

Galerii: rünnak Tuneesia kuurordis

Tuneesia terrori- rünnakus tapeti vähemalt viis briti turisti

Tapatöö Tuneesias: vähemalt 37 turisti tapeti rünnakus hotellile

Islamiräärmuslased ründasid ka Kuveidis: enesetaputerrorist tappis mošees 13 inimest

Ilves terrorirünnakutest: me ei lase end hirmutada ja oleme tugevad

Kuidas valida tulevast eriala

Apteeker annab nõu: nipid, kuidas ohutult päevitada

Kas investeerida regulaarselt või suurem summa korraga?

Populaarsed higistamisvastased tooted ei lõhna ega määri riideid

Erivajadustega inimene – hinnatud töötaja

Jaanipäeva suurim liiklushuligaan oli roolijoodikust põgeneja Teet

Jaak Joala oleks täna saanud 65-aastaseks

Eesti on valmis vastu võtma 84 kuni 156 põgenikku

Kuidas mõjutavad Venemaa sanktsioonid Eesti šokolaaditootjat?

PÄEVAINTERVJUUMartin Hurt: praegustest sammudest Kremli heidutamiseks ei piisa

Peruus laviini alla jäänud Eesti alpinistide surnukehad leiti üles

Swedbank tõstab järsult osade teenuste hindu

Itaalia peaminister ülemkogul: kui see on teie idee Euroopast, võite te selle endale jätta

Islamiriigi rünnak Prantsusmaal: tehasest leiti peata surnukeha, üks ründaja on vahistatud

Sarkozy rünnakust: see on sõjakuulutus tsivilisatsioonile

Prantsuse peaminister: islamiterrorism tabas taas Prantsusmaad

Prantsuse ja Tuneesia presidendid väljendasid solidaarsust

Laevaehitaja selgitab: Vormsi uuel parvlaeval pole midagi viga

Video: arvatavasti kõige kohutavam suvetöö

Tuli võttis kõik, mis võtta andis

Demograafiline «auk» ja kõrgharidusreform toovad kõrgkoolidesse uued õppekavad
Celia Kuningas-Saagpakk: Lampedusast ja mitte ainult
Vene relvajõudude lennuk rikkus Eesti õhupiiri
Voronja galeriis võtab ilmet kadunud kunstnike näitus
Kuidas mõjutavad emotsioonid ajataju?
Eesti epeenaiskond kaotas Euroopa mängude finaalis kindlalt Rumeeniale
Lugeja küsib: kuidas toimida, kui firma ei maksa proovipäeva eest tasu?
Venemaal peljatakse alade annekteerimist Hiina poolt
Tsipras lükkas tagasi ELi ultimaatumid ja väljapressimise
Viking Lotto suurvõidu saanud paar korraldab pulmapeo
Suvesoe peaks järgmisel nädalal kohale jõudma, aga kauaks seda jagub?
TIIT TUUMALU INTERVJUU RIHO UNDIGAUus Oscari vääriline Eesti film?
Ujuja Daniel Zaitsev pääses Euroopa mängudel finaali
Rein Taaramäe ja Tanel Kangert stardivad Tour de France'il
Interaktiivne graafik: jälgi Wimbledoni tenniseturniiri põnevamaid sündmuseid
Estonian Air riigiabi otsuse venimisest: peata olek pärsib meie arengut
BBC lõi pretsedendi ja asus avaldama «unustatud» artiklite nimekirja
Cameron usub uutesse suhetesse Euroopa Liiduga
AC Milan sõlmis tähtsa lepingu
Soomlased ootavad rekordilist mustikasaaki
Eesti käsipallikoondis võitis Moldovat ja lõpetas turniiri viiendana
Eesti-Leedu-Läti koostööfilm sai Euroopast 120 000 eurot
Rõivaesemed, mis teevad su tervisele salamisi liiga
Eesti esindaja Kelly Kangur sai «Miss Eurasia 2015» viie parema hulka
Ateena Akropoli Parthenoni templis oli iidsel ajal hoiul miljoneid hõbemünte
Postimehe otseülekanne Seto Folgilt: Antti Paalaneni kontsert
Politsei keset igapäevast peredraamat: poeg lõi taldrikuga pähe, naine ründas elukaaslast noaga
Harju maakohus kuulutas välja Aviesi pankroti
Aita leida 34-aastast naist, keda politsei otsib seoses raske kuriteoga
Riigikantselei uuring: enam kui pooled Eesti elanikud pole pagulaste vastu
Ligi: õpetajate palgakasv on eelarveprioriteet
Dijsselbloem: Kreeka kokkulepe peab sündima laupäeval
Peruu fännitar tõmbab Copa Americal tähelepanu
Nädalavahetuse ilm tuleb kuivapoolne
Afganistani rekordiline oopiumisaak tõi turule odava heroiini
Lugeja küsib: miks on teenustasu juhilubade tervisetõendi eest nii kallis?
Järvevana teel on liiklus avariide tõttu häiritud
Mart Kalm: puupealinn?!

Putin Obamale: Ukrainas pole Vene vägesid
Venemaa ei laienda toiduembargot veinile ega ka kondiitritoodetele
Varoufakis: Kreekale esitati teostamatu kokkulepe
Kristina Kallas pagulaste vastuvõtmisest: Eestil on hea stardipositsioon teiste vigadest õppimiseks
Kontaveit kohtub Wimbledoni avaringis endise maailma esireketiga
BBC: Chelsea väravavaht on suurkonkurendi juures arstlikus kontrollis
Suri Vene ekspeaminister Jevgeni Primakov
Meeta Haldre, Anna Haava elu lõpuperioodi hea sõbranna, on lahkunud
Cofidis Gert Jõeäärt Touril osalevasse meeskonda ei valinud
Uus tehnoloogia võimaldab trükkida riietele venivat elektroonikat
Läti luureülem: Vene eriteenistuste aktiivsus on «keskmisel tasemel»
Reinsalu tahab pankrotikuriteod rangema järelevalve alla võtta
FIFA presidendi saaga: Blatter pole enda sõnul tagasi astunud
Eestlased ei tahagi suvel puhata
Harjutus, mis tõstab enesehinnangut
Mida jääb laps sinust mäletama?
Projektikirjutajatel algasid töised ajad
YLE: Soome kavatseb kahe aasta jooksul vastu võtta 800 põgenikku
Kiiev: Ukraina sõjaväe positsioone rünnati ööpäevaga 86 korral
Süüria Hasakeh' linnast põgenes IS-i pealetungi eest 60 000 inimest
Hiina koorid esitavad popkooripeol Eesti hitte
Tasuta supp täidab raskustes laste kõhtu
Jaanus Karilaid: Jevgeni, ära kuula oravate paraadkõnesid!
Video: Neptuuni suurune planeet, mis näeb välja nagu komeet
Vaid 16 protsenti Facebooki tehnoloogiatöötajatest on naised
Nalja kah: lõbus kuulutus pakub müügiks ausat mehekoobast
Postimehe reedene käik turule: maasikatest - oma või võõras, vahet pole
Suur galerii: Ugalas läheb kõik müügiks
Vahva nipp: istuta taim sidrunikesta sisse
Kui palju peab trenni tegema pärast suviste lemmiksöökide tarbimist?
Kümme märki, et sul on kadestusväärne kallim
Bangkok – metropol ja agul
Postimehe katse: milline piim peab kõige kauem külmkapis avatult vastu
Kui kaua pärast «kõlblik kuni» kuupäeva võib piima veel söögiks kasutada?
Valitsusjuhid jõudsid kokkuleppele: kohustuslikke pagulaskvoote ei tule
Rõivas jäi Eesti pakkumise osas ebamääraseks
EL: vabatahtlik pagulaste ümberpaigutamise skeem saab paika juuliks
Jaapanlannade seksisümbol on isane gorilla

Siberi igikeltsast leiti mammuti asemel mumifitseerunud koer

Video: USA järves tekkis 2,4 meetrise diameetriga veekeeris

Lumi Tartust puges Hiinaski tundliku naha vahele

Venemaal peljatakse alade annekteerimist Hiina poolt

Järgmisel nädalal peaks soe kohale jõudma

Lugejad soovitavad: paik, mida Euroopas kindlasti vaadata

Katse: milline piim peab kõige kauem külmkapis avatult vastu

Lauljatar Merlyn Uusküla kehakaalu langetamise saladus

Uku Suviste imeiluss suhe hakkab lõppema?

Tanel Padarile korraldati kooli lõpetamise puhul üllatuspidu

Galerii:Vaata pilte meeleolukast Sõru Jazz festivalist Hiiumaal!

Galerii:«Le Grande Spektaakel» peol pidutseti Olümpose Jumalannadega

Segasummasuvila stiil omas kodus

Peenrapiirded

Mida teha, kui laps ei kuula mind?

On need hormoonid või on mõni fuuria minu sisse elama asunud?

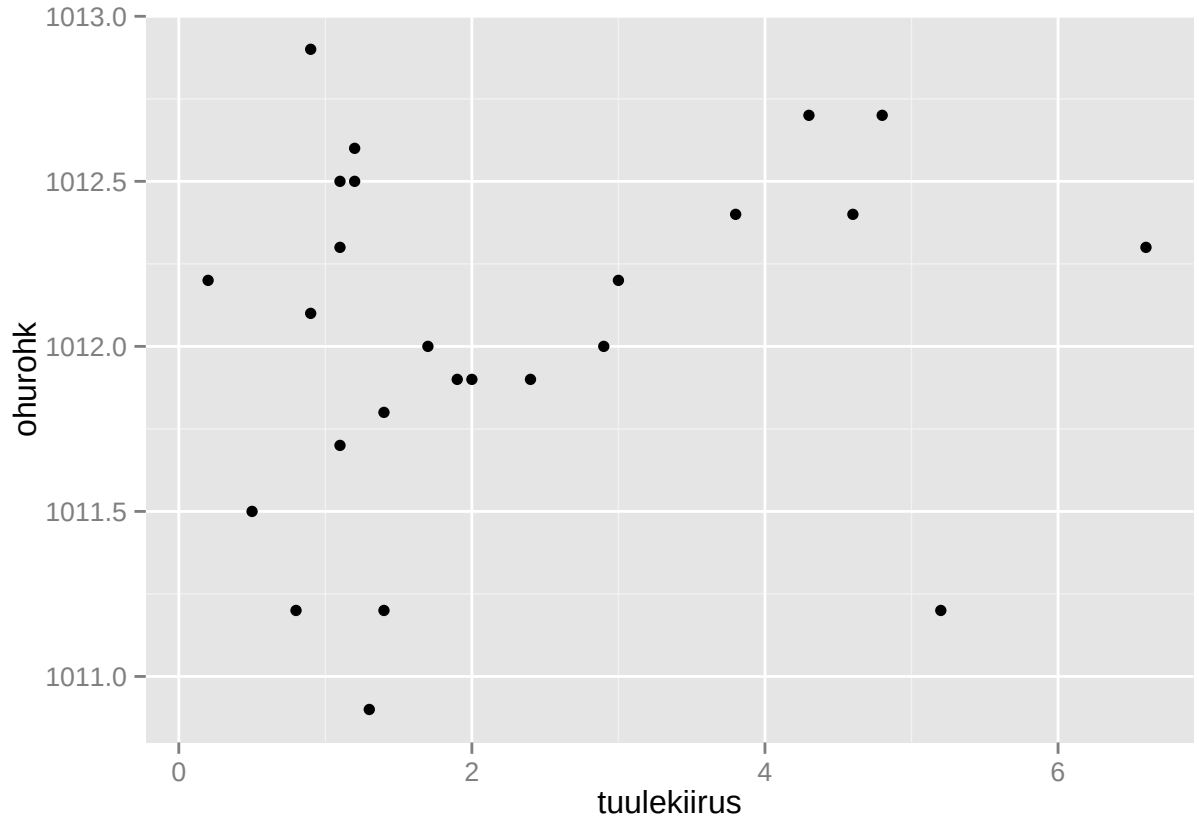
Ülesanne 7

Juhised:

- Riigi Ilmateenistus pakub värskeid ilmaandmeid XML faili kujul.
- Meie tegeleme Eesti vaatlusandmete XML failiga.
- Saa XML failist kätte iga ilmajaama õhurõhk.
- Saa XML failist kätte iga ilmajaama tuule kiirus.
- Tee neist õhurõhu ja tuule kiiruse scatterplot.

```
url="http://www.ilmateenistus.ee/ilma_andmed/xml/observations.php"
page = html(url)

ohurohk=page %>%
    html_nodes("airpressure")%>%
    html_text()

tuulekiirus=page %>%
    html_nodes("windspeed") %>%
    html_text()

nimi=page %>%
    html_nodes("name")%>%
    html_text()

#teeme numericuks
```

```
ohurohk=as.numeric(as.character(ohurohk))
tuulekiirus=as.numeric(as.character(tuulekiirus))
ilm=data.frame(nimi, tuulekiirus, ohurohk)
#teeme graafiku
library(ggplot2)
ggplot(ilm, aes(x=tuulekiirus, y=ohurohk))+
    geom_point()
```

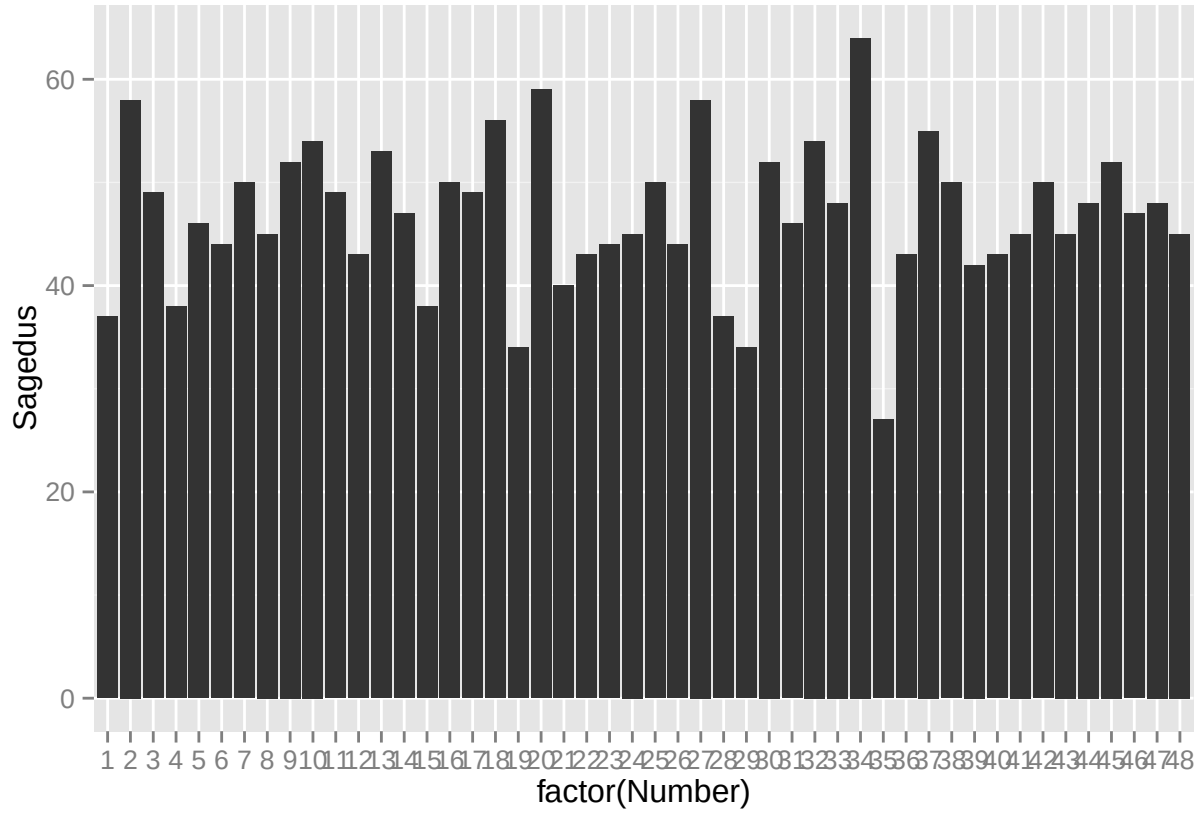


Ülesanne 8

Eesti Loto veebilehel on toodud statistika loositud pallide sagedusest. Eralda vastav tabel, kus veergudes on tunnused number, sagedus ja sagedus protsentides. selectorgadget veab sind siin alt ning kergem on lähtekoodi inspekteerida brauseris olevate tööriistadega (Chrome's vajuta Ctrl + Shift + I või tee parem klikk ja vajuta inspekteeri elementi). Visualiseeri saadud andmetabelit. Tee näiteks tulpdiagramm, kus x-teljel on arvud 1-48 ning y-telg tähistab sagedust.

```
html_source ="https://www.eestiloto.ee/osi/stats.do?lastDraws=250&gameCode=11&sort=frq0&action=searchNu
#lingi osas tegin ise uue päringu ja võtsin siis lingi, algne peksis segast
page = html(html_source)
loto=page %>%
    html_nodes("table") %>%
    html_table(fill=T)

numbrid=as.data.frame(loto[5])
```

```
#graafik
library(ggplot2)
ggplot(numbrid, aes(x=factor(Number), y=Sagedus))+
    geom_bar(stat="identity")
```



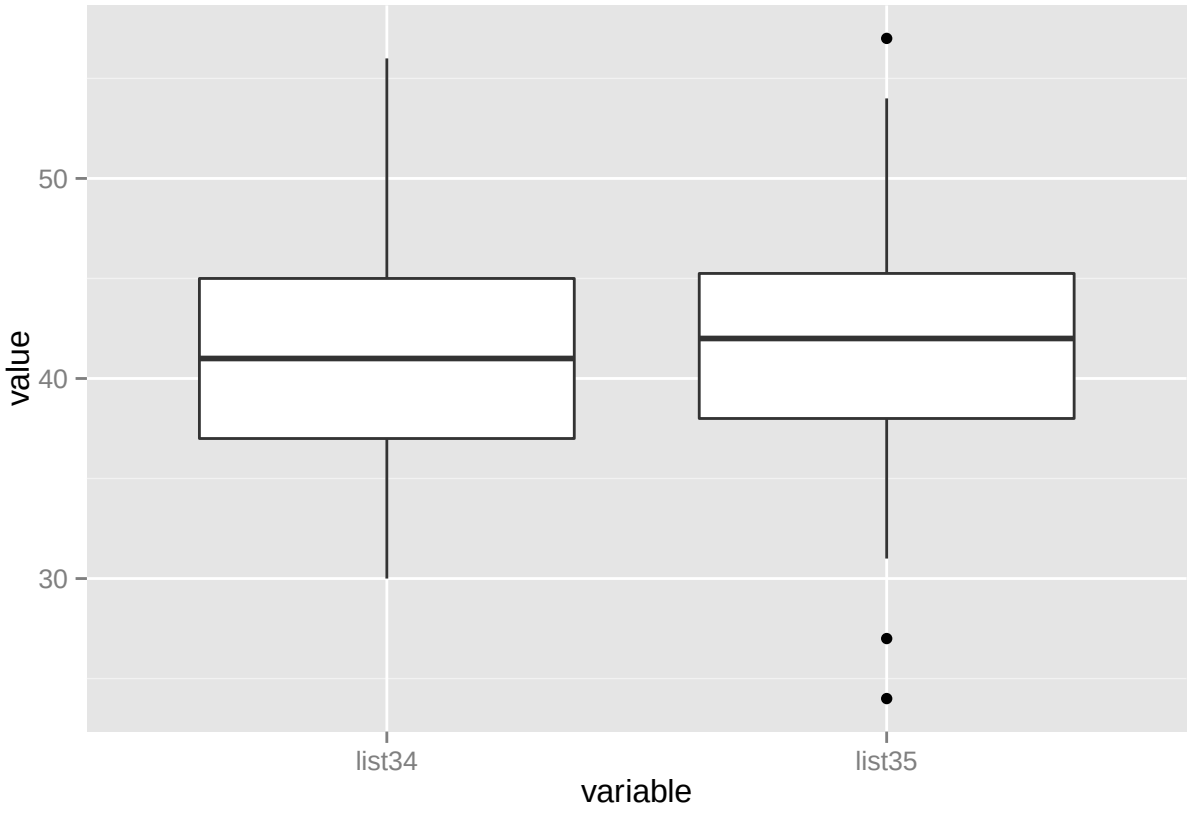
(2 boonuspunkti + lisaboonuspunkt) Viimase 250 loosiga on pall 35 tulnud 28 korral, pall 34 aga 59 korral. Uuri, kas on alust arvata, et Viking Lotto süsteem on kallutatud. Selleks mõtle välja, kuidas seda kontrollida (näiteks võid kasutada simulatsioonidel põhinevat lähenemist). Selgita lühidalt oma lähenemist ja raporteeri, millise tulemuse said. Lisaboonuspunkti saamiseks visualiseeri seda tulemust.

```
#teeme simulatsiooni, iga 250 loosi kohta võtame 34 ja 35 esinemise sageduse
#kordame 100 korda
library(reshape2)
tulem=data.frame(c(1:8))
names(tulem)="järjekord"
list34=list()
list35=list()
j=1
for (j in 1:100)  {

    tulem=data.frame(c(1:8))
    names(tulem)="järjekord"

        for (i in 1:250)
        {
    tulem[,i+1]=sample(1:48, 8, replace=F)
```

```
    names(tulem)[i+1] <- paste("iter", i, sep = "")
        }

    tulem_melt=melt(tulem, id=c("järjekord"))
    #arvutame iga iteratsioonis iga numbri sageduse
    tulem_sagedus=data.frame(table(tulem_melt$value, tulem_melt$variable))
    #hoiame alles ainult 34 ja 35 sagedused, kuna need huvitavad
    tulem_vaja=subset(tulem_sagedus, Var1%in% c(34,35))

    list34[j]=sum(tulem_vaja$Freq[tulem_vaja[,1]==34])
    list35[j]=sum(tulem_vaja$Freq[tulem_vaja[,1]==35])
    j=j+1
}

#teeme dataframeiks listid ja numericuks et ggplottida
simulatsioonid=as.data.frame(rbind(list34, list35))
simulatsioonid=as.data.frame(t(simulatsioonid)) #transpose
simulatsioonid$list34=as.numeric(simulatsioonid$list34)
simulatsioonid$list35=as.numeric(simulatsioonid$list35)
simulatsioonid_melt=melt(simulatsioonid)
```

```
## No id variables; using all as measure variables
```

```
simulatsioonid_melt$value=as.numeric(simulatsioonid_melt$value)
#siit on hästi näha, mis on 34 ja 35 esinemissageduste jaotus 250 numrbite
#võtmise korral lotos
ggplot(simulatsioonid_melt, aes(x=variable, y=value))+
    geom_boxplot()
```


50
40
30
value
list34
list35
variable