Tallinna Ülikool

Digitehnoloogiate instituut

Praktikumimaterjalide koostamine õpikule R for Data Science

Bakalaureusetöö

Autor: Henri Ruut

Juhendaja: Jaagup Kippar

Autor: ..... ,, ... ,, 2017
Juhendaja: ..... ,, ... ,, 2017
Instituudi direktor: ..... ,, ... ,, 2017

Tallinn 2017

Autorideklaratsioon

Deklareerin, et käesolev bakalaureusetöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

.... (kuupäev) (autor)

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina _____ (sünnikuupäev: _____)

(*autori nimi*)

1. annan Tallinna Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   _____

   _____

   _____

   (*lõputöö pealkiri*)

   mille juhendaja on _____,

   (*juhendaja nimi*)

   säilitamiseks ja üldsusele kättesaadavaks tegemiseks Tallinna Ülikooli Akadeemilise Raamatukogu repositooriumis.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tallinnas/Haapsalus/Rakveres/Helsingis, _____

(digitaalne) allkiri ja kuupäev

Sisukord

Sissejuhatus

Tänu meeletule tehnoloogia arengule räägitakse igapäevaselt aina rohkem statistikast ja andmeanalüüsist. Tänapäeval kus kõigil on nutitelefonid ja kogu informatsiooni vahetus toimub internetis toob kaasa suurel hulgalt andmeid, mille põhjal saab väga palju uut informatsiooni erinevates valdkondades. IBM hinnangul 90% andmeid mis on kogutud on tekitatud viimase kahe aasta vältel.

Uuringute järgi on Ameerika kõige parem töökoht olnud viimased kaks aastat andmeteadlane ja nõudlus teadlaste järgi kasvab igapäevaselt (Glassdoor, kuupäev puudub).

Selleks, et algajad ja mitte programmeerimist õppijad saaksid samuti tegeleda andmeanalüüsiga on loodud mitmeid kasutajasõbralike programme ja keskkondi millest üks neist on R ja Rstudio. 2016 aasta lõpus tuli välja uus õpik R for Data Science mis on mõeldud just algajatele ja seda õpikut kasutatakse ka TLÜ magistriõppes**.** Raamatu eesmärk on anda ülevaade kuidas toimub üldiselt andmeteadus, mis on selle eesmärgid ja ülesanded ja samuti õpetab programeerimist R-is. Raamat on ülesehitatud Tidyverse teegile mis on arendatud välja raamatu autori poolt.

Tidyverse on teek mis ühendab omavahel teised pakid mis on mõeldud andmeteaduseks ja moodustab ühtse terviku, mis on mugav alustavatele andmeteadlastele.

Töö teema valik tugineb autori huvist andmeteaduse vastu ja samuti juhendaja soovile. Õppetamise raames on juheldaja täheldanud, et õpik on tihtipeale keeruline ja kõige rohkem on õpilaste seas tekitanud segadust raamatu viies peatükk mis räägib andmete manipuleerimisest.

Seega on töö eesmärgiks koostada täiendav ja toetav praktikumimaterjal R for Data Science õpiku viiendale peatükile, mis aitab õpilasel saada paremat ülevaadet kuidas andmete manipuleerimine käib. Lisaks luua näiteülesandeid, mis oleksid sarnased õpiku omadele, et õpilane oleks suuteline lahendama raamatus pakutuid ülesandeid iseseisvalt.

1 R-programmeerimiskeel

Selles peatükis annab autor ülevaate R-ist, mis on R ja mis on tema tugevused, eelised teiste analüütiliste programmeerimiskeelte ees. Samuti antakse ülevaade mis on üldiselt andmeteadus ja mis on andmeteaduse eesmärkideks.

1.1 Mis on R?

R on programmeerimiskeel, mida kasutatakse peamiselt statistika ja graafiliste analüüside loomiseks. Tänu lisapakettide lisamise võimalusele on tegemist väga paindliku keelega ja annab eelise paljude teiste statistiliste programmeerimiskeelte ees. Üheks R-i tugevuseks on hästi disainitud graafilised joonised. Hästi on lahendatud kasutaja ja programmi koostöö, kus kasutajal on kogu kontroll joonistuse üle ja programm suudab sellest luua hea visuaalse pildi (The R Foundation, kuupäev puudub).

1.2 Mis on andmeteadus?

Andmeteadus on väga kiiresti kasvav valdkond mis ühendab omavahel programmeerimise, statistika, matemaatika ja tihtipeale ka ärinduse. Andmeteadus haarab enda alla väga paljusid valdkondi. Teadust tehakse näiteks sotsioloogias, majanduses, meditsiinis kui ka ärinduses (Frank Lo, 2016).

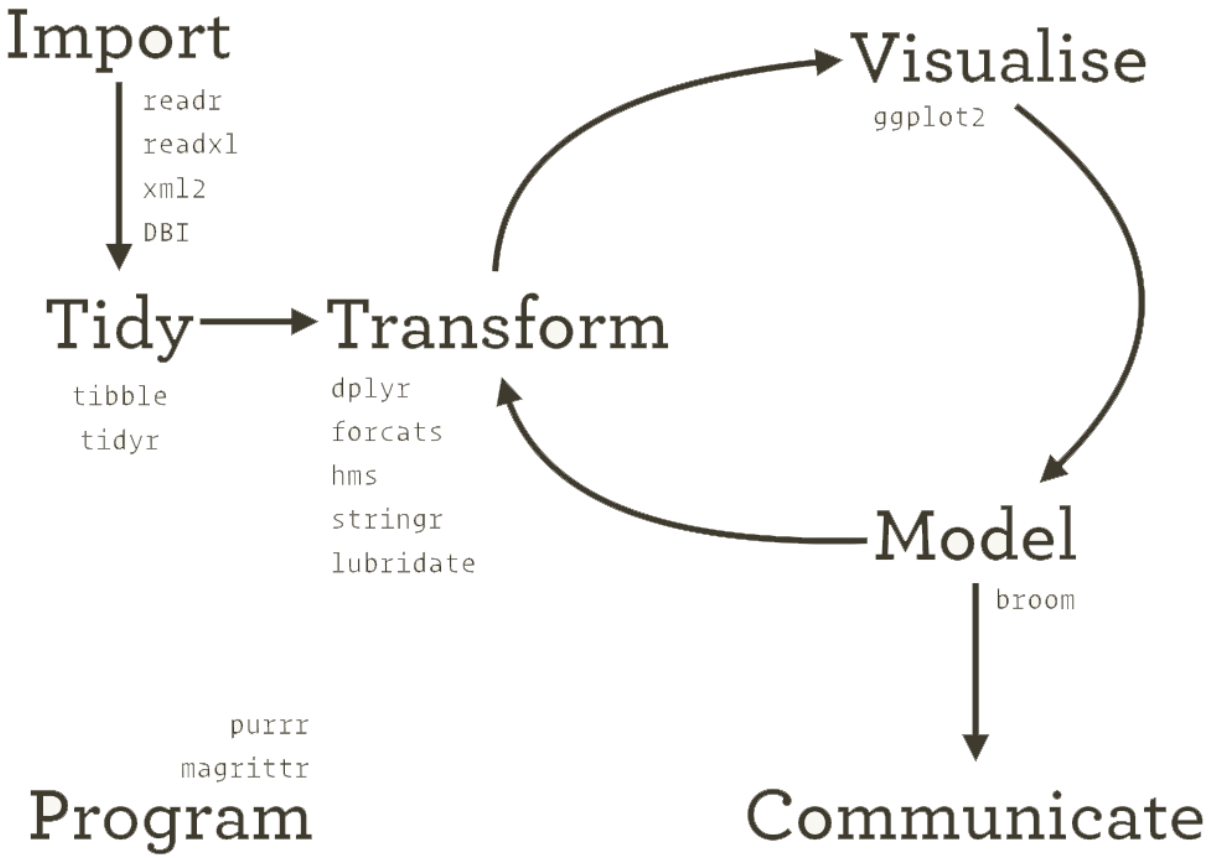
Andmeteaduse eesmärk on luua automatiseeritud meetodeid suurte andmete analüüsimiseks ja selle põhjal uue informatsiooni ja teadmiste loomiseks. Näiteks Netflix jälgib enda kasutajate vaatamisharjumusi, et aru saada missugused filmid ja seriaalid on kasutajate hulgas kõige populaarsemad ja sellepõhjal otsustatakse missugust sisu missugustele kasutajatele toota.

IBM’i hinnangul ligi 90% andmeid tänapäeva maailmas on loodud viimase kahe aasta jooksul ja arvatakse, et aastal 2020 on tekkinud maailmas 50 korda rohkem andmeid kui aastal 2011 (UC Berkeley School of Information, New York University, kuupäev puudub).

2 Tidyverse

Tidyverse on teek, mille eesmärgiks on lahendada keerulisi andmeanalüüsi probleeme kombineerides lihtsaid ja ühtseid tükke. Tegemist on teegiga mis ühendab mitu teist teeki omavahel, et luua nii-öelda ühtne programm. Tidyverse sisaldab järgmisi pakke:

- Readr – andmete sisselugemiseks
- Tidyr – andmete korrastamiseks
- Tibble – kaasaegne andmeteraamistik
- Dplyr – andmete manipuleerimiseks
- Ggplot2 – andmete graafiliseks visualiseerimiseks
- Purrr – funktsionaalseks programmeerimiseks



Joonis 1. Andmeanalüüsi ja Tidyverse pakettide seos.

Joonisel on näidatud skeem, mis järjekorras protsess välja peaks nägema ja millise tegevuse juures millist teeki kasutatakse.

Kõige esimene tegevus on andmete **import** R-i, mis enamasti tähendab andmestiku lisamist mõnest failist või andmebaasist. Kui andmed on süsteemi saadud tuleb need andmed **korrastada**. See on tähtis selleks, et andmestik oleks arusaadav ja loetav, et ei peaks kirjutama keerulisi funktsioone, et kätte saada vajaliku informatsiooni. Peale andmete korrastamist järgneb andmete **muutmine** ja kitsendamine, mille eesmärgiks välja valida vajalikud andmed ja juurde lisada vajalikud tulbad mis on tähtsad analüüsiks. **Visualiseerimine** on tähtis andmestiku ülevaatuseks. Visuaalselt saab inimene parema ülevaate olemas olevast informatsioonist, leida tulemusi mida ei oodatud või püstitada uusi, täiendavaid küsimusi andmestiku kohta. **Mudelid** on täiendav tööriist visualiseerimiseks. Mudelite abil vastatakse püstitatud küsimustele ja tehakse kokkuvõte andmestikust. **Kommunikatsioon** on andmeteaduse viimane komponent ja ühtlasi ka kõige tähtsam. Olgu visuaalid ja mudelid nii ilusad ja head kui tahes, aga kui ei osata neid lahti seletada siis pole nendest mingit kasu. (Grolemund, G., Wickham, H., 2016)

2.1 Tidyverse teek Readr

Teegi ülesandeks on pakkuda võimalikult mugavat ja lihtsalt andmete sisselugemist. See pakk on väga paindlik erinevate faililainendite suhtes, näiteks on võimalik lugeda sisse nii csv, tsv kui ka fwf tüüpi andmestike.

Täpseks andmete sisselugemiseks tuleb kombineerida kaks osa. Esiteks funktsioon mis analüüsib kogu faili ja teiseks tulpade spetsifikatsioonid. Tulpade spetsifikatsioon kirjeldab iga tulba kohta, milline on tulba andmetüüp, kas näiteks täisarv või komakohaga arv. Enamik kordadest pole tegevus vajalik kuna readr teek suudab ise välja lugeda mis andmetüübiga tulbas on tegemist. Reard toetab seitset erinevat failiformaati ja iga tüübi jaoks on vastav sisselugemise funktsioon:

- Read_csv() – komaga eraldatud csv fail
- Read_tsv() – tab-iga eraldatud fail
- Read_delim() – üldine millegagi eraldatud fail (funktsioonis tuleb parameetrina määrata eraldustähis)
- Read_fwf() – fikseeritud laiusega fail (funktsioonis tuleb parameetrina määrata iga välja pikkus)
- Read_table() – failid, mis on eraldatud tühikumärgiga

- Read_log() – log tüüpi failide sisselugemiseks

Näide:

```
mtcars <- read_csv(readr_example("mtcars.csv"))
#> Parsed with column specification:
#> cols(
#>   mpg = col_double(),
#>   cyl = col_integer(),
#>   disp = col_double(),
#>   hp = col_integer(),
#>   drat = col_double(),
#>   wt = col_double(),
#>   qsec = col_double(),
#>   vs = col_integer(),
#>   am = col_integer(),
#>   gear = col_integer(),
#>   carb = col_integer()
#> )
```

Konsooli kuvatakse kõik tulpade spetsifikatsioonid. See on tähtis selleks, et saada ülevaade, kas teek on õigesti arusaanud andmetüüpidest (Wickham, H, kuupäev puudub).

2.2 Tidyverse teek Tidyr

Antud teegi ülesanne on luua „tidy tüüpi andmestike“. Tidy, eesti keeles koristatud andmed on andmed, kus:

- Iga muutuja on tulbas
- Iga rida on vaadeldav tulba andmetega
- Igas lahtris on väärtus

Joonis 2. Andmete paiknemine andmetabelis(mida tähendab missugune nimetus)(Grolemund, G., Wickham, H., 2016).

Tihti öeldakse, et 80% andmeanalüüsist kulub andmete puhastamiseks ja ettevalmistuseks. Samuti tuleb andmeid koristada pidevalt analüüsi käigus kui tekivad uued probleemid või

tuleb juurde uusi andmeid (Wickham, H, kuupäev puudub).

Näide korrastamata ja korrastatud andmestikust:

Korrastamata

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Korrastatud

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Joonis 3. Näide korrastamata ja korrastatud andmestikust(Supervised Machine Learning in Pega Decisioning Solution using R(Part 2), 2016)

2.3 Tidyverse teek Tibble

Tibble asendab R-i data.frame andmestikutüübi. Tibble tüüpi andmestik jätab andmed mis on vajalikud ja viskab välja ebavajalikud. Tegemist on nii-öelda laisa andmestikuga, mis ei muuda tulpade nimesid, ei muuda ridade andmetüüpe ja samuti muudab andmete konsooli väljakuvamist paremaks. Välja prinditakse alati 10 esimest rida koos tulpade nimede ja tüüpidega. Samuti kuvatakse välja palju on ridu kokku ja näitab mitu rida, tulpa on andmestikus (Wickham, H, kuupäev puudub).

Näide:

```
tibble(x = 1:1000)
#> # A tibble: 1,000 × 1     #Andmestiku suurus
#>       x
#>   <int>                   #Tulba andmete tüüp
#> 1     1
#> 2     2
#> 3     3
#> 4     4
#> # ... with 996 more rows  #Mitu rida veel
```

2.4 Tidyverse teek Dplyr

Teek Dplyr vastutab Tidyverses andmete manipuleerimise eest, mis sisaldab viite põhifunktsiooni millega andmestikust kätte saada andmeid. Selle teegi eeliseks on tema lihtsus ja loogilisus, kuna kõik funktsioonid on samasuguse ülesehitusega ja parameetreid sisestatakse alati samas järjekorras (kõigepealt andmestiku nimi ja seejärel komaga eraldatud

vaja minevad tulpade nimed) (Wickham, H, kuupäev puudub).

Kuna antud teek nõuab kõige rohkem programmeerimisalaseid teadmisi ja R for Data Science õpik on andmete manipuleerimise teema juures algajale üsna keeruline siis on püstitatud lahendatavaks probleemiks selle peatüki toetava praktikumimaterjali loomine. Täpsema seletuse koos koodinäidetega leiab peatükist 3.7 ja samuti autori poolt loodud praktikumimaterjalis.

2.5 Tidyverse teek ggplot2

Teek ggplot2 on vajalik graafiliste kokkuvõtete loomiseks. Kasutaja peab sisestama andmed, missugused on graafilise osa tähtsad parameetrid ja missugust graafilist väljundit kasutatakse. Ülejäänud detailide, jooniste, kujundite ja värvidega tegeleb programm ise (Wickham, H, kuupäev puudub).

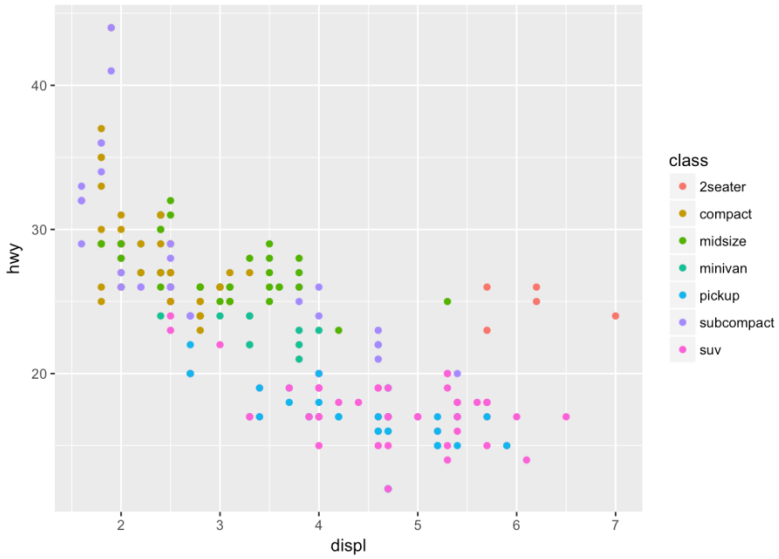
Näide:

```
library(tidyverse)

ggplot(mpg, aes(displ, hwy, colour = class)) + geom_point()
```

Funktsiooni esimeseks parameetriks on andmestik (mpg), teiseks on tähtsad parameetrid millest soovitakse graafiline väljund luua aes(tulpade nimed) ja viimakse geom_point() mis ütleb, et tegemist peab olema hajuvusdiagrammiga.

Väljund:



Joonis 4. Graafiline väljajoonistus R-is ggplot2 teegiga (Wickham, H, kuupäev puudub).

2.6 Tidyverse Teek Purrr

Purrr võimendab R-i programeerimise funktsionaalsust, mille eesmärgiks on erinevad funktsioonid omavahel ühendada ja kuna R-i keelt on vahel üsna ebamugav lugeda siis on võimalik täna Purrr-ile kombineerida funktsioone hästi loetavas koodikeeles. Eraldi tükid seotakse omavahel kasutades Pipe (%>%) operaatorit. Tänu sellele lüheneb kordades koodimaht ja samuti muudab see kergemaks vigade otsimise (Wickham, H, kuupäev puudub).

Näide:

```
library(tidyverse)

df <- mtcars
df %>%
  select(1:3) %>%
  filter(mpg > 20, cyl == 6)
```

andmetabelist mtcars selekteerime 3 esimest tulpa ja seejärel filtreerime nendest välja tulemused mille mpg väärtus on suurem kui 20 ja cyl väärtus on kuus.

Väljund:

```
   mpg cyl disp
1 21.0   6  160
2 21.0   6  160
3 21.4   6  258
```

2.7 Tidyverse kirjapildi erinevus traditsioonilisest R-ist

Kuna baas/traditsiooniline R on üsna „omapärane“ programmeerimiskeel ja koosneb väga paljudest erinevatest funktsioonidest on sellega töötamine üsna keerukas protsess. Samuti on R-is võimalus lahendada ühte probleemi mitut erinevat pidi mis lisab oma korda veel kompleksust.

Tidyverse teek on loodud selleks, et muuta koodipilt loetavamaks ja kasutajasõbralikumaks, ehk isegi programmeerimist mitteõppiv inimene peaks kiiresti saama selgeks kuidas käib R-is andmetega ringi käimine.

Tidyverse’s andmete manipuleerimise eest vastutab „dplyr“ mille funktsioonid aitavad kergesti lahendada mitmeid ülesandeid, näiteks tulba selekteerimine, ümbernimetamine, andmete muutmine ja filtreerimine. Tänu lihtsusele on tegemist R ühe populaarsema lisateegiga. (Muenchen, B, 2017)

Ridade filtreerimine

```
# base R
crime.ny.2005 <- crime.by.state[crime.by.state$Year==2005 &
                                crime.by.state$State=="New York", ]

# dplyr
crime.ny.2005 <- filter(crime.by.state, State=="New York", Year==2005)
```

Näites on näha, et traditsioonilise R-i puhul on filtreerimine üsna keeruline. Andmestiku nimi on crimes.by.state, nurksulge kasutatakse sellest, et öelda programmile mis tulbast mis tulbani vaja andmeid võtta. $ tähistab tulba nimetust. Dplyr’ga on filtreerimine kordades lihtsam ja koodi on kergem lugeda. Funktsiooni esimene parameeter on alati andmestik ja kõik järgmised komaga eraldatud parameetrid on loogikatehted.

Tulpade korrastamine

```
# base R
crime.ny.2005 <- crime.ny.2005[order(crime.ny.2005$Count, decreasing=TRUE), ]

# dplyr
crime.ny.2005 <- arrange(crime.ny.2005, desc(Count))
```

Baas R-i puhul võtakse andmestikust kõik Count read ja pannakse need kahanevasse järjekorda. Dplyr’i puhul on näha samat tulemust, aga koodipilt on kordades puhtam ja arusaadavam.

Tulpade selekteerimine

```
# base R
crime.ny.2005 <- crime.ny.2005[, c("Type.of.Crime", "Count")]

# dplyr
crime.ny.2005 <- select(crime.ny.2005, Type.of.Crime, Count)
```

Uue tulba lisamine

```
# base R
crime.ny.2005$Proportion <- crime.ny.2005$Count /sum(crime.ny.2005$Count)

# dplyr
crime.ny.2005 <- mutate(crime.ny.2005, Proportion=Count/sum(Count))
```

Kogu programm

```
# base R
crime.by.state <- read.csv("CrimeStatebyState.csv")
crime.ny.2005 <- crime.by.state[crime.by.state$Year==2005 &
                                crime.by.state$State=="New York",
                                c("Type.of.Crime", "Count")]
crime.ny.2005 <- crime.ny.2005[order(crime.ny.2005$Count, decreasing=TRUE), ]
crime.ny.2005$Proportion <- crime.ny.2005$Count / sum(crime.ny.2005$Count)
summary1 <- aggregate(Count ~ Type.of.Crime, data=crime.ny.2005, FUN=sum)
summary2 <- aggregate(Count ~ Type.of.Crime, data=crime.ny.2005, FUN=length)
final <- merge(summary1, summary2, by="Type.of.Crime")


# dplyr
crime.by.state <- read.csv("CrimeStatebyState.csv")
final <- crime.by.state %>%
         filter(State=="New York", Year==2005) %>%
         arrange(desc(Count)) %>%
         select(Type.of.Crime, Count) %>%
         mutate(Proportion=Count/sum(Count)) %>%
         group_by(Type.of.Crime) %>%
         summarise(num.types = n(), counts = sum(Count))
```

Kokkuvõtvalt võib öelda, et nii koodipildi kui ka arusaadavuse poolepealt on dplyr algajale ja mitte igapäevaselt programmeerijale igati vastuvõetavam. Funktsioonid on lihtsad, lühikesed ja loogilise ülesehitusega. Samuti ajakulu poolest võttis antud näide baas R-ist 50% vähem aega. (Fischetti, T, 2014)

3 Õppematerjalide ülevaade

Selles peatükis räägib autor headest olevasolevatest õppematerjalidest, mida autor soovitab läbi töötada, et saada parem ülevaade kuidas teha andmeteadust R-is. Kõik õppematerjalid on mõeldud algajatele.

3.1 R for Data Science

Autor(id)	Hadley Wickham, Garrett Grolemund
Tüüp	E-õppematerjal, raamat
Aadress	http://r4ds.had.co.nz
Sihtrühm	Algajad

Raamatu eesmärgiks on anda lugejale võimalikult hea ülevaade andmeteadusest ja kuidas kasutada R-i selle loomiseks. Raamatus käsitletakse kõiki nii-öelda andmeteaduse tsükkli osi alustades andmete õigele kujule viimisest ja lõpetades mudelite loomisega. Raamat koosneb viiest suuremast peatükist mis seletavad lahti iga andmeteaduse etapi. Lisaks on materjalis

väga palju koodinäiteid koos seletusega ja ülesannetega. Raamat on mõeldud inimestele kellel puudub varasem programmeerimise kogemus, kuid siiski on juhendaja täheldanud, et algajatel tekib selle õpikuga raskusi. Tihti jäävad õpilased hätta andmete muutmise peatüki juures kus ülesanded on üsna keerulised. Sellepärast näeb juhendaja täiendava praktikumimaterjali vajadust.

3.2 R Programming for Data Science

Autor(id)	Roger D. Peng
Tüüp	Raamat
Aadress	https://leanpub.com/rprogramming
Sihtrühm	Algajad

Antud raamat räägib R programmeerimiskeele põhitõdedest ja andmestike manipuleerimisest. Tegemist on rohkem tehnilise õpperaamatuga kus keskendutakse rohkem programmeerimisele kui andmeteaduse õpetamisele. Autor valis selle raamatu ülevaateks sellepärast, et seal käsitletakse väga detailselt ’dplyr’ teeki mis vastutab Tidyverses andmete manipuleerimise eest. Raamatus on palju näiteid kõikide funktsioonide kohta ja samuti on iga koodirea kohta väga detailne seletus.

3.3 Working efficiently with large datasets

Autor(id)	Gergana, John, Francesca, Sandra and Isla
Tüüp	E-õppematerjal
Aadress	https://ourcodingclub.github.io/2017/03/20/seecc.html
Sihtrühm	Algajad

Tegemist on e-õppematerjaliga/kursusega, mille eesmärgiks on anda kiire ülevaade kuidas toimub andmetega käitumine R-is. Autor valis selle õppematerjali selleks kuna seal kasutusel olevad teegid on samad mis Tidyverses ja see kursus täiendab R for Data Science raamatut. Kursuse raames tuuakse välja erinevate lähenemiste head ja halvad küljed, räägitakse erinevate funktsioonide ajakulust ja võtakse põhjalikult läbi graafiline andmete kuvamine.

4 Praktikumimaterjali loomine

Bakalaureusetöö eesmärgiks on luua täiendav praktikumimaterjal, mis toetab R for Data Science õpikut. Materjali eesmärgiks on anda laiem ülevaade õpiku viiendast peatükist kuna selle koodinäited koos ülesannetega on programmeerimist mitteoskavale inimesele keerulised.

Eesmärgi saavutamiseks on autor läbi töötanud õpiku ja samuti teised õppematerjalid, mis on sarnased õpiku viiendale peatükile. Seejärel koostas autor enda materjali ja lasi seda testida sihtrühma kuuluva isiku peal.

4.1 Praktikumimaterjali vajadus

Praktikumimaterjali loomise vajadusest lähtus autor juhendaja soovist. Nimelt R for Data Science õpikut kavatsetakse kasutada loodava haridusanalüütika magistrimooduli juures. Kuna eelnevalt on õppejõud samuti kasutanud õpikut ja selle tulemusena on täheldanud, et raskusi valmistab õpilaste jaoks viies peatükk siis otsustas autor koos juhendajaga luua täiendavat materjali antud õpiku täiendamiseks.

4.2 Ülesehitus

Kõigepealt räägitakse materjalis täiendavalt andmete manipuleerimisest ja siis kirjeldatakse täpsemalt antud teeki, mis selle tegevuse eest vastutab. Koodinäited on tehtud võimalikult lihtsad ja arusaadavad, et materjali kasutav inimene saaks peale vaadates aru mis antud funktsiooniga on tehtud. Samuti on iga koodinäite kohta koodi kirjeldav jutt ja sarnane ülesanne, et õpilane saaks õpitut kinnitada.

4.3 Õpiväljundid

Õpilane, kes on läbi töötanud autori poolt tehtud praktikumimaterjali omab laiemaid teadmisi dplyr teegist ja samuti andmete sisselugemisest. Õpilane oskab lahendada iseseisvalt õpiku viienda peatüki ülesandeid ja samuti luua keerulisemaid statistilisi analüüse omale huvi pakkuvast andmestikust.

Kokkuvõte

Bakalaureusetöös tutvustati R-i programmeerimiskeelt, anti ülevaade andmeteadusest ja uuriti olemasolevaid õppematerjale. Autor töötas läbi kõik õppematerjalid, et sellepõhjal koostada enda praktikumimaterjal.

Bakalaureusetöö peamiseks eesmärgiks oli koostada täiendav praktikumimaterjal R for Data Science õpikule. Kuna õpik on mõeldud programmeerimist mitteoskavatele inimestele siis kavatsetakse kasutada õpikut TLÜ magistriõppes haridusanalüütika mooduli juures. Õppejõud on õpiku kasutamise käigus märganud, et raamatul on hea teoreetiline osa kuid harjutusülesanded ja koodinäited on õpitu kinnitamiseks liiga keerulised või liiga lihtsakoeliselt seletatud. Kõige rohkem tekkis õpilastel probleeme viienda peatüki läbitöötamisel. Selle tõttu sai koostatud täiendav praktikumimaterjal R for Data Science õpiku viiendale peatükile.

Materjali koostamisel lähtus autor peamiselt R for Data Science õpiku formaadist ja ülesehitusest. Kuna praktikumimaterjal on mõeldud algajatele ja mitte programmeerimist oskavatele inimestele siis andis autor enda poolt koostatud praktikumimaterjali testimiseks ka sihtrühma kuuluvale isikule, kust sai autor tagasisidet materjali parendamiseks.

Bakalaureusetöö käigus valmis täiendav praktikumimaterjal õpiku viiendale peatükile. Praktikumimaterjal annab täiendavat informatsiooni dplyr teegi kõikidest funktsioonidest ja lisaks on igale funktsioonile tehtud täiendavad koodinäited koos ülesannetega.

Bakalaureusetööd tehes õppis autor palju uusi teadmisi andmeteadusest, tidyverse teegist ja praktikumimaterjal koostamisest.

Summary

In the bachelor's thesis, introduction to the R programming language and overview to the data science was given. Already existing literature about R programming language was also researched. Author worked through all the study material, to make his own material for studying the programming language

Bachelor's thesis main objective was to make additional practicum material for the textbook called "R for Data Science". Because the textbook is meant for the people who don't know how to program, then the intention of this textbook is to use it for TLÜ master's degree students, among the educational analysis module. While using the textbook "R for Data Science", lecturers have noticed that the book has good theoretical part, but exercises and code examples are too complicated to put the theoretical knowledge to work. Other option is that the theoretical part does not provide enough knowledge for the book's practical part. Students had most trouble when working through the fifth chapter. That is why an additional study material was made for the fifth chapter of the book "R for Data Science".

When writing the study material, author relied on the book "R for Data Science" format and structure. Because the study material is meant for beginners and for people who do not know how to program, author tested the study material on a student who belongs to the target audience. Author got feedback from the student to further improve the study material

While writing the study material, additional practicum material was made for the fifth chapter of the book "R for Data Science". The study material gives additional information about all dplyr library functions and for every function there are additional code examples with exercises.

While writing the bachelor's thesis, author learnt a lot about data science, tidyverse library and composing study material.

Kasutatud kirjandus

Fischetti, T. (2014). *How dplyr replaced my most common R idioms*. Kasutamise kuupäev 26.aprill 2017.a, allikas http://www.onthelambda.com/2014/02/10/how-dplyr-replaced-my-most-common-r-idioms/

Frank Lo. (2016). *What is Data Science?*. Kasutamise kuupäev 28.aprill 2017.a, allikas https://datajobs.com/what-is-data-science

Glassdoor. (kuupäev puudub). *50 Best Jobs in America*. Kasutamise kuupäev: 28.aprill 2017.a, allikas https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Grolemund, G., Wickham, H. (2016) *R for Data Science* . O'Reilly. http://r4ds.had.co.nz/

Muenchen, B. (2017). *The Tidyverse Curse*. Kasutamise kuupäev 30.aprill 2017.a, allikas http://r4stats.com/2017/03/23/the-tidyverse-curse/

New York University. (kuupäev puudub). *What is Data Science?*. Kasutamise kuupäev 01.mai 2017.a, allikas http://datascience.nyu.edu/what-is-data-science/

The R Foundation. (kuupäev puudub) *What is R?*. Kasutamise kuupäev: 21.aprill 2017.a, allikas https://www.r-project.org/about.html

UC Berkeley School of Information. (kuupäev puudub) *What is Data Science?*. Kasutamise kuupäev 01.mai 2017.a, allikas https://datascience.berkeley.edu/about/what-is-data-science/

Wickham, H. (kuupäev puudub) *The Tidyverse*. Kasutamise kuupäev 20.aprill 2017.a, allikas http://tidyverse.org/

Lisad

Lisad 1. Praktikumimaterjal

Tallinna Ülikool

Digitehnoloogiate instituut

Praktikumimaterjalid õpikule R for Data Science

Praktikumimaterjal

Autor: Henri Ruut

Tallinn 2017

Sisukord

1 Andmete manipuleerimine

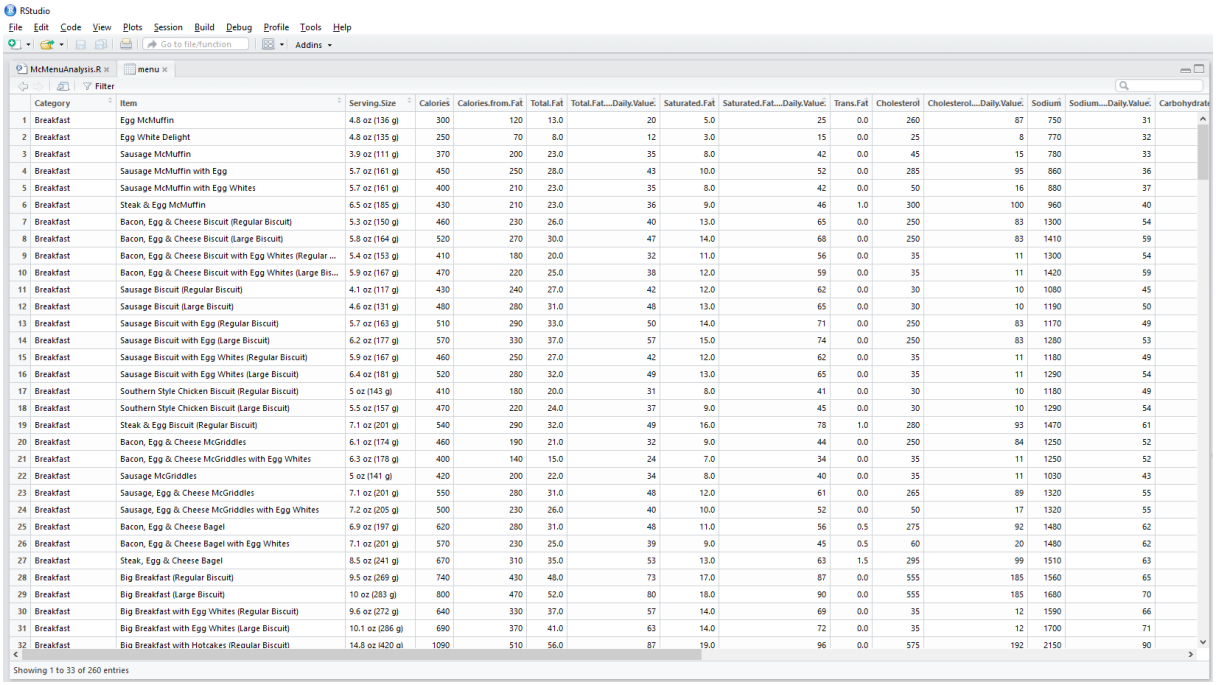
Selleks, et luua mõistliku visuaalset andmestiku tuleb tihtipeale andmeid manipuleerida. Luua uusi tulpasid, järjestada ümber olemas olevaid tulpasid või neid omavahel grupeerida. Selleks on R'is olemas tidyverse teek, mis muudab andmete töötlemise lihtsamaks. Samuti on R Studios sisseehitatud mugav funktsioonide õpetus. Selleks, et saada mingi teatud funktsiooni kohta rohkem informatsiooni tuleb konsooli kirjutada ?funktsiooni_nimi. Näiteks ?select(). Seejärel kuvatakse välja funktsiooni dokumentatsioon koos lihtsamate näidetega.

2 Andmete sisselugemine ja ülevaade

Esmalt tuleb lisada programmi teek, mis lubab „Tibble" tüüpi andmestike ümber järjestada ja nende andmeid manipuleerida. Selleks tuleb konsooli kaudu installida „pakk" mis sisaldab vajaminevat teeki: `install.packages("tidyverse")`. Seejärel tuleb sisse lugeda andmestik. Andmestiku on võimalik lugeda erinevates formaatides nagu näiteks: XML ja CSV. Kuna antud hetkel on tegu CSV formaadis sisendfailiga tuleb kasutada funktsiooni `read.csv( faili_nimi, header = TRUE ).` Header parameeter ütleb, et failis on ka tulpade pealkirjad olemas. Et andmestik oleks Tibble tüüpi tuleb rakendada funktsiooni `tbl_df().` Samuti on R'is võimalus lugeda andmeid sisse veebilehtedelt. Selleks tuleb kasutada funktsiooni `url().` Näiteandmestiku sisselugemine käib järgmiselt:

```
library(tidyverse) // tidyverse teek
menu <- tbl_df(read.csv(url("http://www.tlu.ee/~ruut/mcmenu/menu.csv"), header = TRUE))
View(menu)
```

Nüüd on olemas muutuja nimega menu. Tegemist on andmetabeliga millega töötlemist alustatakse. Andmeid saab visuaalselt vaadata konsooli või koodi kirjutades `View( tabeli_nimi ).`



	Category	Item	Serving.Size	Calories	Calories.from.Fat	Total.Fat	Total.Fat....Daily.Value.	Saturated.Fat	Saturated.Fat....Daily.Value.	Trans.Fat	Cholesterol	Cholesterol....Daily.Value.	Sodium	Sodium....Daily.Value.	Carbohydrate
1	Breakfast	Egg McMuffin	4.8 oz (136 g)	300	120	13.0	20	5.0	25	0.0	260	87	750	31	
2	Breakfast	Egg White Delight	4.8 oz (135 g)	250	70	8.0	12	3.0	15	0.0	25	8	770	32	
3	Breakfast	Sausage McMuffin	3.9 oz (111 g)	370	200	23.0	35	8.0	42	0.0	45	15	780	33	
4	Breakfast	Sausage McMuffin with Egg	5.7 oz (161 g)	450	250	28.0	43	10.0	52	0.0	285	95	860	36	
5	Breakfast	Sausage McMuffin with Egg Whites	5.7 oz (161 g)	400	210	23.0	35	8.0	42	0.0	50	16	880	37	
6	Breakfast	Steak & Egg McMuffin	6.5 oz (185 g)	430	210	23.0	36	9.0	46	1.0	300	100	960	40	
7	Breakfast	Bacon, Egg & Cheese Biscuit (Regular Biscuit)	5.3 oz (150 g)	460	230	26.0	40	13.0	65	0.0	250	83	1300	54	
8	Breakfast	Bacon, Egg & Cheese Biscuit (Large Biscuit)	5.8 oz (164 g)	520	270	30.0	47	14.0	68	0.0	250	83	1410	59	
9	Breakfast	Bacon, Egg & Cheese Biscuit with Egg Whites (Regular ...	5.4 oz (153 g)	410	180	20.0	32	11.0	56	0.0	35	11	1300	54	
10	Breakfast	Bacon, Egg & Cheese Biscuit with Egg Whites (Large Bis...	5.9 oz (167 g)	470	220	25.0	38	12.0	59	0.0	35	11	1420	59	
11	Breakfast	Sausage Biscuit (Regular Biscuit)	4.1 oz (117 g)	430	240	27.0	42	12.0	62	0.0	30	10	1080	45	
12	Breakfast	Sausage Biscuit (Large Biscuit)	4.6 oz (131 g)	480	280	31.0	48	13.0	65	0.0	30	10	1190	50	
13	Breakfast	Sausage Biscuit with Egg (Regular Biscuit)	5.7 oz (163 g)	510	290	33.0	50	14.0	71	0.0	250	83	1170	49	
14	Breakfast	Sausage Biscuit with Egg (Large Biscuit)	6.2 oz (177 g)	570	330	37.0	57	15.0	74	0.0	250	83	1280	53	
15	Breakfast	Sausage Biscuit with Egg Whites (Regular Biscuit)	5.9 oz (167 g)	460	250	27.0	42	12.0	62	0.0	35	11	1180	49	
16	Breakfast	Sausage Biscuit with Egg Whites (Large Biscuit)	6.4 oz (181 g)	520	280	32.0	49	13.0	65	0.0	35	11	1290	54	
17	Breakfast	Southern Style Chicken Biscuit (Regular Biscuit)	5 oz (143 g)	410	180	20.0	31	8.0	41	0.0	30	10	1180	49	
18	Breakfast	Southern Style Chicken Biscuit (Large Biscuit)	5.5 oz (157 g)	470	220	24.0	37	9.0	45	0.0	30	10	1290	54	
19	Breakfast	Steak & Egg Biscuit (Regular Biscuit)	7.1 oz (201 g)	540	290	32.0	49	16.0	78	1.0	280	93	1470	61	
20	Breakfast	Bacon, Egg & Cheese McGriddles	6.1 oz (174 g)	460	190	21.0	32	9.0	44	0.0	250	84	1250	52	
21	Breakfast	Bacon, Egg & Cheese McGriddles with Egg Whites	6.3 oz (178 g)	400	140	15.0	24	7.0	34	0.0	35	11	1250	52	
22	Breakfast	Sausage McGriddles	5 oz (141 g)	420	200	22.0	34	8.0	40	0.0	35	11	1030	43	
23	Breakfast	Sausage, Egg & Cheese McGriddles	7.1 oz (201 g)	550	280	31.0	48	12.0	61	0.0	265	89	1320	55	
24	Breakfast	Sausage, Egg & Cheese McGriddles with Egg Whites	7.2 oz (205 g)	500	230	26.0	40	10.0	52	0.0	50	17	1320	55	
25	Breakfast	Bacon, Egg & Cheese Bagel	6.9 oz (197 g)	620	280	31.0	48	11.0	56	0.5	275	92	1480	62	
26	Breakfast	Bacon, Egg & Cheese Bagel with Egg Whites	7.1 oz (201 g)	570	230	25.0	39	9.0	45	0.5	60	20	1480	62	
27	Breakfast	Steak, Egg & Cheese Bagel	8.5 oz (241 g)	670	310	35.0	53	13.0	63	1.5	295	99	1510	63	
28	Breakfast	Big Breakfast (Regular Biscuit)	9.5 oz (269 g)	740	430	48.0	73	17.0	87	0.0	555	185	1560	65	
29	Breakfast	Big Breakfast (Large Biscuit)	10 oz (283 g)	800	470	52.0	80	18.0	90	0.0	555	185	1680	70	
30	Breakfast	Big Breakfast with Egg Whites (Regular Biscuit)	9.6 oz (272 g)	640	330	37.0	57	14.0	69	0.0	35	12	1590	66	
31	Breakfast	Big Breakfast with Egg Whites (Large Biscuit)	10.1 oz (286 g)	690	370	41.0	63	14.0	72	0.0	35	12	1700	71	
32	Breakfast	Big Breakfast with Hotcakes (Regular Biscuit)	14.8 oz (420 g)	1090	510	56.0	87	19.0	96	0.0	575	192	2150	90	

Joonis 5. View() funktsiooniga välja kuvatud visuaalne tabel Rstudios.

Sedasi on hea ülevaade antud andmestikust, kus näeme tulpi ja ridu selgelt. Antud näite puhul on tegemist McDonald's menüüga kus on kõikide toitude/jookide toitainete sisaldused. Ühel real on märgitud toidu kategooria, nimetus, kaal, kalorid jne. Hea andmestiku puhul on tulpade pealkirjad informatiivsed ja arusaadavad.

2.1 Dplyr alused

Eelnevalt installitud teegis on olemas viis põhifunktsiooni, mille abil saab lahendada enamus andmete manipuleerimised:

- Filter() – funktsioon mille abil saab välja filtreerida vajaminevad andmed
- Arrange() – funktsioon tulpade ümberjärjestamiseks
- Select() – funktsioon tulpade nimepidi selekteerimiseks
- Mutate() – funktsioon uue tulba loomiseks olemasolevatest tulbaandmetest
- Summarise() – funktsioon tulpade summeerimiseks

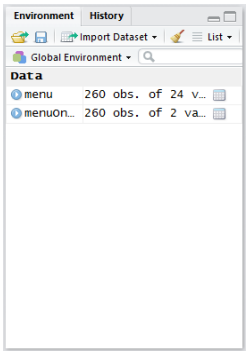
Funktsiooni loogika on lihtne. Funktsiooni nimi ja sulgude sees on argumendid. Esimeseks argumendiks on alati andmestiku nimi ja kõik järgmised parameetrid on andmed mida soovitakse muuta, valida jne.

2.1.1 Tulpade selekteerimine select() funktsiooniga

Tihti on olukord, kus andmestikus on sadu või tuhandeid tulpi ja nende tulpade vahel navigeerimine võib osutuda väga keeruliseks. Selleks on olemas select() funktsioon, mille abil saab välja võtta ainult vajaminevad tulbad.

```
menuOnlyCategory <- select(menu, Category, Item)
> (menuOnlyCategory <- select(menu, Category, Item))
# A tibble: 260 × 2
    Category                                                        Item
      <fctr>                                                      <fctr>
1  Breakfast                                                Egg McMuffin
2  Breakfast                                            Egg White Delight
3  Breakfast                                             Sausage McMuffin
4  Breakfast                                    Sausage McMuffin with Egg
5  Breakfast                           Sausage McMuffin with Egg Whites
6  Breakfast                                        Steak & Egg McMuffin
7  Breakfast                 Bacon, Egg & Cheese Biscuit (Regular Biscuit)
8  Breakfast                   Bacon, Egg & Cheese Biscuit (Large Biscuit)
9  Breakfast Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit)
10 Breakfast   Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)
# ... with 250 more rows
```

Nüüd on salvestatud uus tabel mille nimi on menuOnlyCategory.



Joonis 6. Keskkonda loodud andmestikud

Mitme järjestikuse tulba selekteerimiseks on võimalik kasutada koolonit. Nimelt tulp millest alates on andmeid vaja seejärel koolon ja siis tulba nimi millega selekteerimine peaks lõppema. Näiteks kui on vaja võtta andmestikust kõik read tulbast Category kuni Calories siis näeks päring välja järgmine:

```
> (menuOnlyCalories <- select(menu, Category:Calories))
# A tibble: 260 × 4
    Category                                                    Item   Serving.Size Calories
      <fctr>                                                  <fctr>         <fctr>    <int>
1  Breakfast                                          Egg McMuffin 4.8 oz (136 g)      300
2  Breakfast                                      Egg White Delight 4.8 oz (135 g)      250
3  Breakfast                                       Sausage McMuffin 3.9 oz (111 g)      370
4  Breakfast                              Sausage McMuffin with Egg 5.7 oz (161 g)      450
5  Breakfast                     Sausage McMuffin with Egg Whites 5.7 oz (161 g)      400
6  Breakfast                                  Steak & Egg McMuffin 6.5 oz (185 g)      430
```

Ülesanne:

Tekita uus andmetabel niiet tabelis oleks ainult valitud ainult kategooria, nimetus ja kalorite sisaldus.

2.1.2 Andmestiku filtreerimine filter() funktsiooniga

Filter funktsioon on vajalik, kui on vajalik andmete filtreerimine. Funktsiooni esimeseks argumendiks on andmestik ja järgmisteks argumentideks on tulpade nimed ja loogikatehted nagu näiteks: ja, ei, või. Iga loogikatehe peab vastuseks andma õige, et antud tulba filtreering töötaks. Loogikatehete tähendused:

- y & x – kehtima peavad mõlemad tehted, näiteks kaloreid vähem kui 400 ja kategooria hommikusöök
- y & !x – peab kehtima y ja mitte x, näiteks kaloreid vähem kui 400 ja kategooria ei tohi olla hommikusöök (Calories < 400 & !Category == ’Breakfast’)
- y | x – peab kehtima y või x, näiteks kaloreid vähem kui 400 või rohkem kui 500 (Calories < 400 | Calories > 500)

Näiteks kui soovime andmestikust välja võtta kõik hommikusöögid, mille kalorisisaldus on väiksem kui 400 siis näeb päring välja järgmine:

```
> filter(menuOnlyCalories, Calories < 400 & Category == 'Breakfast')
# A tibble: 8 × 4
   Category                                        Item   Serving.Size Calories
     <fctr>                                      <fctr>         <fctr>    <int>
1 Breakfast                                Egg McMuffin 4.8 oz (136 g)      300
2 Breakfast                            Egg White Delight 4.8 oz (135 g)      250
3 Breakfast                             Sausage McMuffin 3.9 oz (111 g)      370
4 Breakfast                                     Hotcakes 5.3 oz (151 g)      350
5 Breakfast                               Sausage Burrito 3.9 oz (111 g)      300
6 Breakfast                                    Hash Brown    2 oz (56 g)      150
7 Breakfast                          Fruit & Maple Oatmeal 9.6 oz (251 g)      290
8 Breakfast Fruit & Maple Oatmeal without Brown Sugar 9.6 oz (251 g)      260
```

Samuti saab kasutada süntaksit x %in% y. Mis tähendab, et x tulbas kõik y väärtused. Näiteks kui me soovime saada liha –ja kalaroogasid mille kalorisisaldus on suurem kui 100 kalorit ja väiksem kui 1000 kalorit, siis on võimalik filtreerida järgmiselt:

```
> filter(menuOnlyCalories, (Calories > 100 & Calories <1000) , Category %in% c('Beef &
Pork','Chicken & Fish'))
# A tibble: 41 × 4
      Category                                      Item   Serving.Size Calories
        <fctr>                                    <fctr>         <fctr>    <int>
1  Beef & Pork                                 Big Mac 7.4 oz (211 g)      530
2  Beef & Pork               Quarter Pounder with Cheese 7.1 oz (202 g)      520
3  Beef & Pork       Quarter Pounder with Bacon & Cheese   8 oz (227 g)      600
4  Beef & Pork Quarter Pounder with Bacon Habanero Ranch 8.3 oz (235 g)      610
5  Beef & Pork                     Quarter Pounder Deluxe 8.6 oz (244 g)      540
6  Beef & Pork        Double Quarter Pounder with Cheese  10 oz (283 g)      750
7  Beef & Pork                                 Hamburger  3.5 oz (98 g)      240
8  Beef & Pork                               Cheeseburger   4 oz (113 g)      290
9  Beef & Pork                         Double Cheeseburger 5.7 oz (161 g)      430
10 Beef & Pork                      Bacon Clubhouse Burger 9.5 oz (270 g)      720
# ... with 31 more rows
```

Ülesanne:

Filtreeri välja kõik suupisted mille kalorite sisaldus ei ole rohkem kui 400 kalorit.

2.1.3 Tulpade järjestamine arrange() funktsiooniga

Funktsioon `arrange()` töötab samamoodi nagu `filter()` kuhu funktsiooni sisse saadetakse tulba nimed mida soovitakse ümber reastada, kas kasvavas või kahanevas järjekorras. Kui on vaja tulp sorteerida väiksemast suuremaks tuleb kasutada funktsiooni `desc( Tulba_nimi ).` Tuleb meeles pidada, et kõik tühjad väärtused sorteeritakse alati tabeli lõppu. Näiteks kui oleks vaja järjestada kõik menüüs olevad toidud kahanevalt kogu rasvasisalduse järgi siis näeb tegevus välja järgmiselt:

```
menuWithFat <- select(menu, Category, Item, Serving.Size, Calories, Total.Fat)
arrange(menuWithFat, desc(Total.Fat))
```

Kõigepealt tekitame uue tabeli milles on kategooria, nimetus, kaal, kalorite sisaldus ja kogu rasva sisaldus. Seejärel järjestame tulba arrange funktsiooni abil ja et andmed oleksid kahanevalt kasutame `desc()` funktsiooni.

```
> arrange(menuWithFat, desc(Total.Fat))
# A tibble: 260 × 5
         Category                                                      Item    Serving.Size Calories Total.Fat
           <fctr>                                                    <fctr>          <fctr>    <int>     <dbl>
1  Chicken & Fish                              Chicken McNuggets (40 piece) 22.8 oz (646 g)     1880       118
2       Breakfast                 Big Breakfast with Hotcakes (Large Biscuit) 15.3 oz (434 g)     1150        60
3  Chicken & Fish                              Chicken McNuggets (20 piece) 11.4 oz (323 g)      940        59
4       Breakfast               Big Breakfast with Hotcakes (Regular Biscuit) 14.8 oz (420 g)     1090        56
5       Breakfast                                 Big Breakfast (Large Biscuit)   10 oz (283 g)      800        52
6       Breakfast   Big Breakfast with Hotcakes and Egg Whites (Large Biscuit) 15.4 oz (437 g)     1050        50
7       Breakfast                               Big Breakfast (Regular Biscuit)  9.5 oz (269 g)      740        48
8       Breakfast Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit) 14.9 oz (423 g)      990        46
9     Beef & Pork                           Double Quarter Pounder with Cheese   10 oz (283 g)      750        43
10      Breakfast               Big Breakfast with Egg Whites (Large Biscuit) 10.1 oz (286 g)      690        41
# ... with 250 more rows
```

Ülesanne:

Järjestada andmetabeli kalorite sisaldus kasvavas järjekorras.

2.1.4 Uue tulba lisamine mutate() funktsiooniga

Tihti tule ette olukord kus andmestikus oleks vaja luua lisatulpa, mis on kokku pandud mitmest olemasolevast tulbast. Selleks on dplyr'is olemas fuktsioon `mutate().` Näiteks kui on vaja rasva kalorite sisalduse protsenti kogu kaloritest ja sellest luua lisatulp. Protsendi leidmiseks on vaja rasva kalorid korrutada 100'ga ja see järel jagada kogu kaloritega. Programmis näeks tegevus välja järgmiselt:

```
nameAndFat <- select(menu, Category:Calories.from.Fat)
mutate(nameAndFat, total.fat.asPerc = Calories.from.Fat * 100 / Calories)
```

Defineerimue uue tabeli kategooriatest kuni rasva kaloriteni. Seejärel lisame tulba funktsiooni `mutate()` abil. Esimene parameeter on tabeli nimi kuhu uus tulp luua. Teiseks parameetriks on uue tulba nimetus ja selle loogikatehe. Võimalik on lisada ka mitu uut tulpa.

```
mutate(nameAndFat, total.fat.asPerc = Calories.from.Fat * 100 / Calories)
# A tibble: 260 × 6
    Category                                                          Item   Serving.Size Calories Calories.from.Fat total.fat.asPerc
      <fctr>                                                        <fctr>         <fctr>    <int>             <int>            <dbl>
1  Breakfast                                                Egg McMuffin 4.8 oz (136 g)      300               120         40.00000
2  Breakfast                                            Egg White Delight 4.8 oz (135 g)      250                70         28.00000
3  Breakfast                                             Sausage McMuffin 3.9 oz (111 g)      370               200         54.05405
4  Breakfast                                      Sausage McMuffin with Egg 5.7 oz (161 g)      450               250         55.55556
5  Breakfast                                Sausage McMuffin with Egg Whites 5.7 oz (161 g)      400               210         52.50000
6  Breakfast                                          Steak & Egg McMuffin 6.5 oz (185 g)      430               210         48.83721
7  Breakfast                 Bacon, Egg & Cheese Biscuit (Regular Biscuit) 5.3 oz (150 g)      460               230         50.00000
8  Breakfast                   Bacon, Egg & Cheese Biscuit (Large Biscuit) 5.8 oz (164 g)      520               270         51.92308
9  Breakfast Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit) 5.4 oz (153 g)      410               180         43.90244
10 Breakfast   Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit) 5.9 oz (167 g)      470               220         46.80851
# ... with 250 more rows
```

Ülesanne:

Luua tulp, kus on küllastunud rasvade protsent kogu rasvast iga rea kohta,

2.1.5 Tulba kokkuvõtte tegemine summarise() funktsiooniga

Viimane põhifunktsioon on `summarise().` Mille funktsiooni nimetuski ütleb, et tegemist on kokkuvõttega. See funtsioon võtab kokku tabeli tulbad üheks reaks. Näiteks kui on vaja leida kõikide söökide keskmine kalorisisaldus oleks tegevus järgmine:

```
summarise(nameAndFat, mean.calories = mean(Calories, na.rm = TRUE))
```

funktsioon `mean()` abil on võimalik arvutada kõikide sisendväärtuste keskmist. Na.rm = TRUE tähendab seda, et keskmise arvutamisest võetakse välja tühjad väärtused, kuna see võib

mõjutada keskmist.

```
# A tibble: 1 × 1
  mean.calories
          <dbl>
1      368.2692
```

Et muuta antud näide pisut informatiivsemaks kasutame funktsiooni `group_by()`. See funktsioon grupeerib sisse antud tulba. Kui on vaja saada näiteks kõikide kategooriate keskmist keskmist kalorisisaldust siis oleks tegevus järgmine:

```
byCaloriesAndFat <- group_by(nameAndFat, Category)
summarise(byCaloriesAndFat, mean.calories = mean(Calories, na.rm = TRUE))
```

Grupeerime andmestiku kategooriate järgi ja seejärel võtame keskmise kaloritest.

```
> summarise(byCaloriesAndFat, mean.calories = mean(Calories, na.rm = TRUE))
# A tibble: 9 × 2
             Category mean.calories
               <fctr>         <dbl>
1        Beef & Pork      494.0000
2          Beverages      113.7037
3          Breakfast      526.6667
4     Chicken & Fish      552.9630
5       Coffee & Tea      283.8947
6            Desserts      222.1429
7              Salads      270.0000
8 Smoothies & Shakes      531.4286
9     Snacks & Sides      245.7692
```

Ülesanne:

- Leida kõige suurem kalorisisaldus kasutades max() funktsiooni.
- Leida kõige suuremad kalorisisaldused kategooriate kaupa kasutades max() funktsiooni