Kodutöö

Töö eesmärk:

Antud töö eesmärgiks on analüüsida vabalt valitud andmestikust leiduvat infot ning luua antud teema kohta ülesandeid, mille lahendamisel kasutan praktikumis omandatud oskusi. Käsitlen antud töös järgmisi teemasid:

Millise andmestikuga on tegemist?

Andmestik omab ülevaadet olümpiamängudelt (2000-2010) võidetud medalite hulgast ja nende võitjatest, sisaldades infot nii võitja rahvuse kui ka vanuse kohta. Lisaks on toodu välja, milliselt spordialalt on medalid võidetud.

I. Andmete sisselugemine, puuduvad väärtused

Tegemist on Exceliga CSV-formaati salvestatud failiga, mistõttu kasutame käsku ***read.csv()***. Antud juhul on komakoha sümboliks punkt.

- > andmed = read.csv("Olümpia.csv", header = T, sep = ";", dec = ".")

Kuna andmestikus on veergudes ***Gold.Medals***, ***Silver.Medals***, ***Bronze.Medals*** puuduvaid väärtusi ***NA***, siis täidame need järgmiselt:

- > andmed$Silver.Medals[is.na(andmed$Silver.Medals)] = 0
  > andmed$Gold.Medals[is.na(andmed$Gold.Medals)] = 0
  > andmed$Bronze.Medals[is.na(andmed$Bronze.Medals)] = 0

Käsuga str(andmed) saame hea ülevaate, mis muutujatüüpidega on veerud. Võime ka kasutada eraldi ka käsku ***class()*** nt >class(andmed$Age). Puuduvate väärtuste asendamisel muutus esialgne muutujatüüp täisarv (integer) reaalarvuks (numeric) veergudes ***Gold.Medals***, ***Silver.Medals*** *ja* ***Bronze.Medals.*** Vajadusel saab muuta tagasi täisarvuks, kui kasutada käsku as.integer().

II. Esmane andmekirjeldus

Uuritavas andmestikus on:

- Ridasid ja veerge

  > x = data.frame(read = nrow(andmed), veerud = ncol(andmed))
  > x
    read veerud
  1 8618   10

- Veergude nimedeks – > names(andmed)
  - "Athlete"
  - "Age"
  - "Country"
  - "Year"
  - "Closing.Ceremony.Date" "Sport"
  - "Gold.Medals"
  - "Silver.Medals"
  - "Bronze.Medals"
  - "Total.Medals"
- Kirjeldav statistika kõigi OM-l osalenute vanuse kohta:

  > summary(andmed$Age)

  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's

  15.00   23.00   26.00   26.41   29.00   61.00   5

III. Kirjeldav statistika

Leiame 2000. aasta olümpimängudel osalenute hulgast vanima sportlase ja noorima kuldmedalivõitja ujumises:

- > max(andmed$Age[andmed$Year==2000])
- > min(andmed$Age[andmed$Sport == "Swimming" & andmed$Gold.Medals])
  - 47
  - 15

Saime, et 2000. aasta olümpiamängudel oli vanim osaleja 47 aastane ning suveolümpiamängude noorim kuldmedali võitja ujumises oli vaid 15. aastane.

Leiame eraldi 2000 ja 2012 olümpiamängudel osalenute keskmised vanused:

- >keskmised=data.frame(mean(andmed$Age[andmed$Year==2000]),mean(andmed$Age[andmed$Year==2012],na.rm=TRUE))
  > colnames(keskmised)=(c("2000", "2012"))
  > rownames(keskmised)="keskmine"

> keskmised

	2000	2012
keskmine	26.27174	26.23885

Näeme, et 2000. a osalenute keskmine vanus võrreldes 2012. a osalenute kekmisest vanusest ei erine palju: 2000. a osalenute keskmine vanus on vaid veidikene suurem.

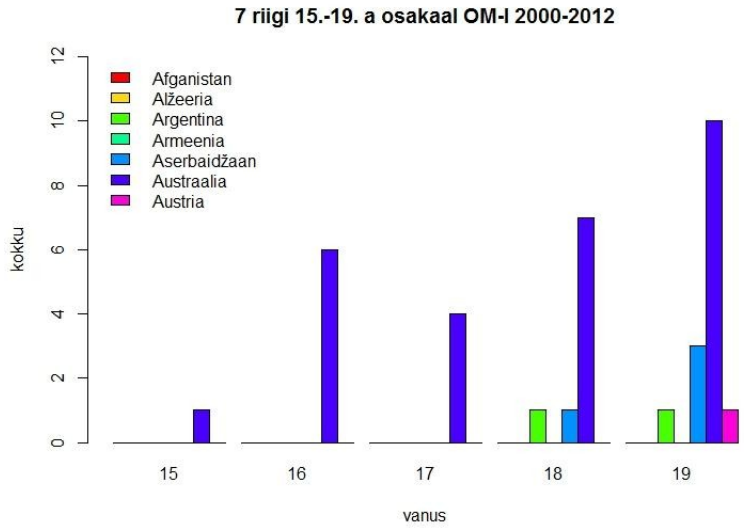
Leiame kui palju sportlasi vanuses 15-19 osales 2000-2012. aasta olümpiamängudel riikide lõikes ja teeme sagedustabeli, kuvades ekraanile vaid esimesed 7 riiki. Kasutame ***table()*** käsku:

- > length(andmed$Age[andmed$Age<20])
  > 510
  Osales 510 sportlast, kes olid nooremad kui 20.
- > v = table(andmed$Country[andmed$Age<20],
  andmed$Age[andmed$Age<20]) # *moodustame sagedustabeli*
  > v[1:7,1:5] # *kuvame tabeli mõõtmetes 5x7*

	15	16	17	18	19
Afghanistan	0	0	0	0	0
Algeria	0	0	0	0	0
Argentina	0	0	0	1	1
Armenia	0	0	0	0	0
Azerbaijan	0	0	0	1	3
Australia	1	6	4	7	10
Austria	0	0	0	0	1

Näeme, et seitsme riigi hulgast on olnud Austraalia osavõtt 15-16 sportlaste seas kõige suurem. Antud sagedustabelis on olümpiamängudelt enim osa võtnuid 19. ja 18. a sportlaste seas.

- >barplot(v[1:7,1:5],ylim = c(0,12),xlab = 'vanus', ylab = 'kokku', beside=T, main='7 riigi 15.-16. a osakaal OM-l 2000-2012',col = rainbow(7))
  legend("topleft", c("Afganistan", "Alžeeria", "Argentina", "Armeenia", "Aserbaidžaan","Austraalia","Austria"), cex=1.0, bty="n", fill=rainbow(7));
  Näeme, et vanuses 19. eluaastat on osavõtt olnud kõige suurem. Lisaks on ülekaalukalt kõige rohkem sportlasi Austraaliast.



Joonis 1 Sagedustabelile vastav tulpdiagramm (kasutatud on barplot() käsku)

Leiame samale sagedustabelile ka jaotustabeli, kasutades ***prop.table()*** käsku:

- > prop.table(v[1:7,1:5])

	15	16	17	18	19
Afghanistan	0.00000000	0.0000000	0.0000000	0.00000000	0.00000000
Algeria	0.00000000	0.0000000	0.0000000	0.00000000	0.00000000
Argentina	0.00000000	0.0000000	0.0000000	0.02857143	0.02857143
Armenia	0.00000000	0.0000000	0.0000000	0.00000000	0.00000000
Azerbaijan	0.00000000	0.0000000	0.0000000	0.02857143	0.08571429
Australia	0.02857143	0.1714286	0.1142857	0.20000000	0.28571429
Austria	0.00000000	0.0000000	0.0000000	0.00000000	0.02857143

Ka jaotustabelist näeme, et rohkem kui 50% vanuses 15-19 osalenutest on rahvuselt austraallased.

IV. Tsükli, funktsiooni, plyr ja ggplot2 kasutamine

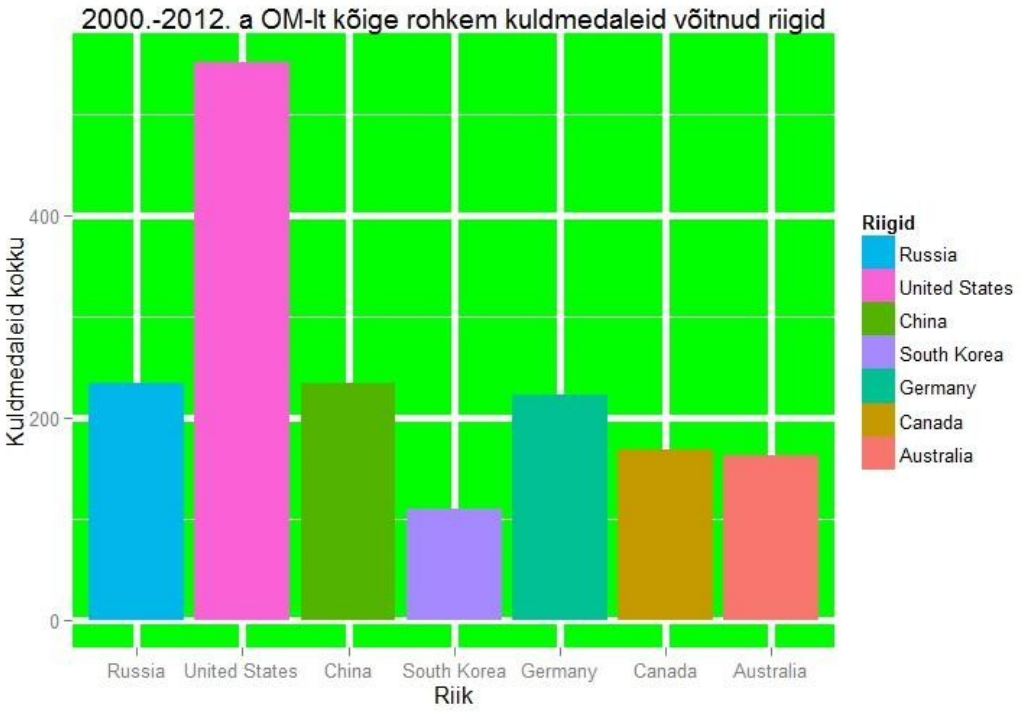
1) Kasutades for tsüklit , kuvame ekraanile OM osalejate arvu riikide lõikes, samuti eraldi riikide vanima ja noorima sportlase, kes vahemikus 2000. - 2012. a OM-lt osa võtnud:

- > for (riik in levels(andmed$Country)){
  if((length(andmed$Athlete[andmed$Country == riik])) == 1){ # *kui osalejate arv on võrdne 1, siis kuvatakse ekraanile*
  print(riik)
  print('On osalenud vaid üks sportlane')
  print("***")
  }
  else{ # *kui ei võrdu 1, siis kuvatakse ekraanile*
  print(riik)

```
                print(length(andmed$Athlete[andmed$Country == riik]))
                print(max(andmed$Age[andmed$Country == riik],na.rm = T))
                print(min(andmed$Age[andmed$Country == riik],na.rm = T))
                print("***")
        }
    }
"Afghanistan"
2               # osalejate arv
25              # vanim osaleja
21              # noorim osaleja
"***"
"Algeria"
8
29
22
"***"
"Argentina"
141
46
18
"***"
```

2) Joonistame tulpdiagrammi enim kuldmedaleid võitnud riikide kohta, kasutades ***qplot()*** ning ka ***plyr()*** käsku:

- ```
  > kuldm = ddply(andmed, "Country", summarize, Kokku = sum(Gold.Medals))
  > qplot(Country, Kokku, data = kuldm, geom = "bar",xlab =
  "Riik",ylab="Kuldmedaleid kokku",main = "2000.-2012. a OM-lt kõige
  rohkem kuldmedaleid võitnud riigid",stat = "identity", fill = Country ) +
  scale_x_discrete(limits = c("Russia","United States","China","South
  Korea","Germany","Canada","Australia"))+
  scale_fill_hue(name = "Riigid", breaks = c("Russia","United
  States","China","South Korea","Germany","Canada","Australia"))
  last_plot() + theme(panel.background = element_rect(fill = "green"),
  panel.grid.major = element_line(size = 2, colour = "white"),
  panel.border = element_blank())
  ```
  ```

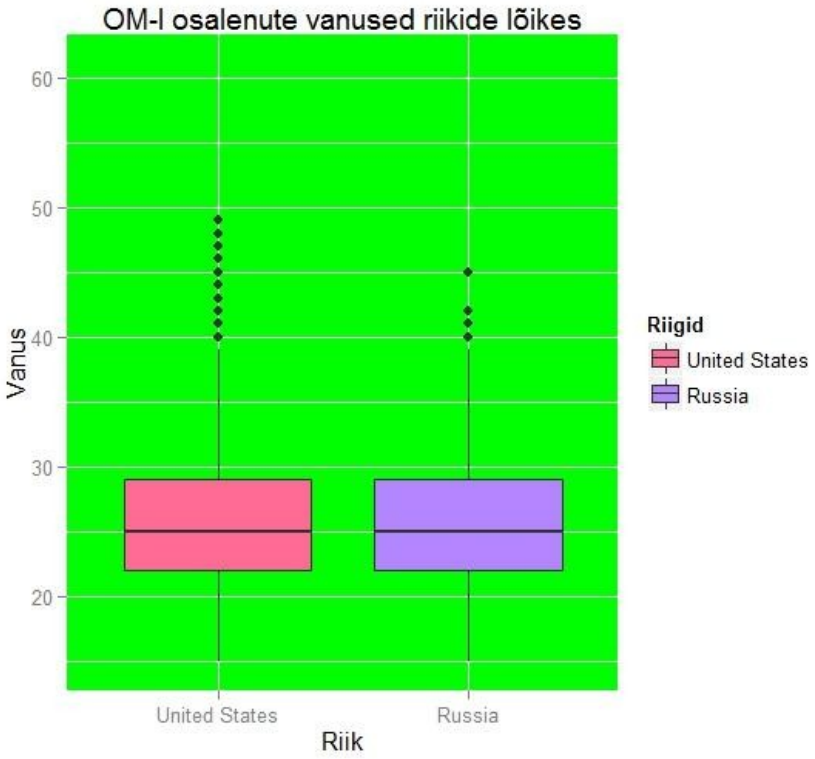


Joonis 2 Tulpdiagramm kõige rohkem kuldmedaleid OM-lt võitnud riikide kohta

Joonisel 2 näeme, et aastatel 2000 - 2012 toimunud OM-lt on kõige rohkem kuldmedaleid võitnud Ameerika Ühendriigid.

3) Joonistame karpdiagrammi võrdlemaks OM-l osalenud Ameerika Ühendriikide ja Venemaa sportlaste vanuseid:

- > qplot(Country,Age, data=andmed ,geom="boxplot", fill=Country,xlab = "Riik", ylab = "Vanus",main="OM-l osalenute vanused riikide lõikes") + scale_x_discrete(limits = c("United States","Russia")) + theme(panel.background = element_rect(fill = "green")) + scale_fill_hue(name = "Riigid", breaks = c("United States", "Russia"))



Joonis 3 Karpdiagramm OM-l (2000-2012) osalenud Venemaa ja USA sportlaste vanuste jaotuse kohta

- Võime kontrollida, kas karpdiagrammil olevad andmed on tõesed, selleks kasutame kirjeldava statistika käske ***median(), quantile():***

```
> quantile(andmed$Age[andmed$Country=="United States"], p = c(0.25, 0.5,0.75), na.rm=T) # USA sportlased

> quantile(andmed$Age[andmed$Country=="Russia"], p = c(0.25, 0.5,0.75), na.rm=T) # Venemaa sportlased
```

USA			Venemaa		
25%	50%	75%	25%	50%	75%
22	25	29	22	25	29

Näeme tõepoolest, et joonisel 3 olevad andmed on tõesed, sest nii ülemine kui ka alumine ning mediaan (50% tähistab mediaani) kattub arvutuste teel saadud tulemustega. Lisaks saame väita, et USA kui ka Venemaa sportlaste vanuste mediaan ning ülemine kui ka alumine kvartiil ühtivad. Siiski tuginedes joonisele saame öelda, et vanuste suurim väärtus on antud juhul erinev.

- Kirjutame lihtsa funktsiooni, mis arvutab kvantiilide haarde, seega:

```
> haare = function(x,x1,x2) {
        h = quantile(x, p=x1) - quantile(x, p=x2)
```

```
        return (print(c("Haare on:",h)))
}
haare(andmed$Age[andmed$Country=="United States"], 0.75,0.25) #
```

katsetame seda USA sportlaste puhul

```
                    75% - 25%
Kvantiilide haare on:       7
```

V. Pikk ja lai formaat *cast(), melt()*

- Teisendame andmestikust 50 rida ja 6-9 veergu pikka formaati ***melt()*** käsku kasutades:

```
> s=melt(andmed[1:50, 6:9])
```

```
      variable value
1  Gold.Medals     8
2  Gold.Medals     6
3  Gold.Medals     4
4  Gold.Medals     1
5  Gold.Medals     2
6  Gold.Medals     1
      ...
```

Variable veeru alla koonduvad ***Gold.Medals, Silver.Medals, Bronze.Medals.***

- Seejärel kasutame sõnetöötlust ja lisame juurde veel ühe veeru nimega tyyp:

```
> s$tyyp = as.character(s$variable)
> s$tyyp = str_replace(s$tyyp, ".Medals", "") # asendab sõna ".Medals" jättes
```

alles kas Gold, Silver või Bronze

- Teeme graafiku:

Uurime kui palju pronksmedaleid on võidetud spordialade lõikes:

```
> s2 = ddply(s[s$tyyp == "Bronze", ], "Sport", summarize, med_saak = value)
> qplot(Sport, med_saak, data = s2,xlab="Spordiala",ylab="Medaleid kokku",
geom= "bar",fill = Sport, stat = "identity")
last_plot() + theme(axis.text.x = element_text(angle = 45,size=7))
```