COVID-19 plitimas valstybėse

Laboratorinis darbas

Antanas Užpelkis, Matas Amšiejus

Vilniaus Universitetas

2020-05-24

Duomenys yra paimti iš EU Open Data Portal apie koronaviruso plitimą skirtingose pasaulio valstybėse. Duomenys kaupti nuo 2019 gruodžio 31 iki dabar, tačiau mes failą parsisiuntėme gegužės 10 dieną. Duomenys yra tokie: data(pilna), diena, mėnuo, metai, užsikrėtimų tą dieną skaičius vienetais, mirusių nuo koronaviruso skaičius tą dieną (vienetais), valstybė, valstybės geo ID, valstybės kodas, pupuliacija 2018 metų duomenimis, žemynas (pastaba: Š ir P Amerikos čia imamos kaip vienas žemynas Amerika). Duomenų eilučių yra 16741 (mūsų faile) ir kasdien pildoma.

**Laiko eilučių duomenys**

**1**

Analizavome Lietuvoje nustatytų Covid 19 atvejų kitimą laike nuo kovo mėnesio (nuo laiko kada virusas pradėjo plisti Lietuvoje). Pirma nuskaitėme duomenis iš .csv failo:

duom<-read.csv('D:/<...> /covid.csv)

Pasirenkame tik Lietuvos susirgimų ir datos duomenis iš failo:

duomenys<-subset(duom,countriesAndTerritories=="Lithuania”, select=c(cases, dataRep))

Pakeičiame lentelės stulpelių vardus:

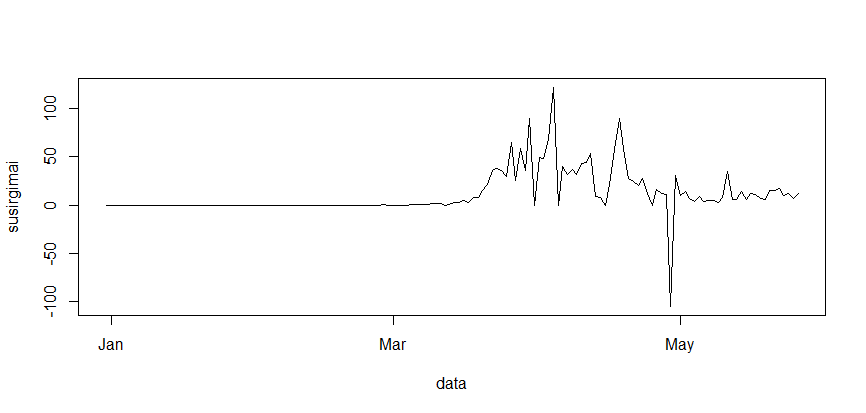
names(duomenys)<-c("susirgimai", "data")

Konvertuojame nuskaitytus datos duomenis dataRep į „Data“ tipo duomenis:

duomenys$data1 = as.Date(duomenys$data,"%d/%m/%Y")

Sukuriame grafiką:

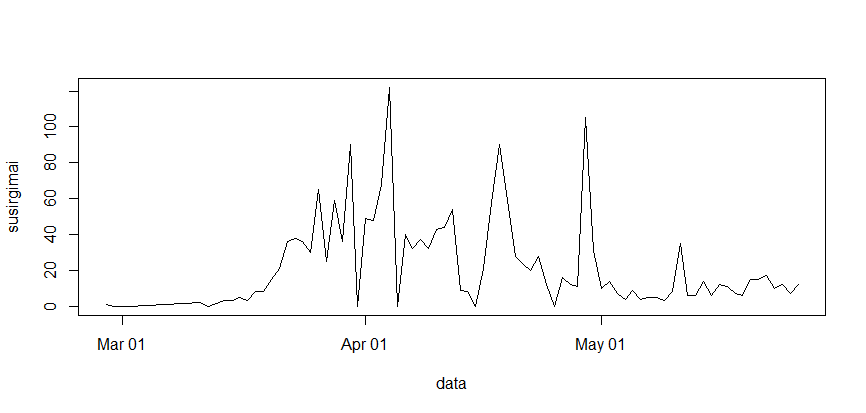
plot(duomenys$data1, duomenys$susirgimai, type = “l”, xlab = “data”, ylab = “susirgimai”)

Matome kelias problemas: neigiamus susirgimus ir daugybę 0 duomenų pradžioje. Norint pašalinti nulius reikia atsekti kada OFICIALIAI prasidėjo virusas Lietuvoje. Paaiškėjo, kad vasario 28 d. Vadinasi, imame tik duomenis, kurie prasideda nuo šios datos:

duomenys<-subset(duomenys, duomenys$data1 >= "2020-02-28", select = c(susirgimai, data1))

Pasidomėjus paaiškėjo, kad neigiama reikšmė yra teisingo dydžio, tiesiog su neigiamu ženklu. Tą sutvarkome panaudojus modulio funkciją:

duomenys$susirgimai<-abs(duomenys$susirgimai)

Gauname grafiką:

**2**

Randame duomenų didžiausio sergamumo indeksą:

imax<-match(max(duomenys$susirgimai), duomenys$susirgimai)

Gavus indeksą galima sužinoti kurią dieną ir kiek buvo daugiausiai susirgimų:

duomenys$data1[imax] 122

duomenys$susirgimai[imax] "2020-04-04"

Randame duomenų mažiausio sergamumo indeksą:

imin<-match(min(duomenys$susirgimai), duomenys$susirgimai)

Tada analogiškai:

duomenys$data1[imin] 0

duomenys$susirgimai[imin] "2020-04-25"

Randame susirgimų vidurkį:

vidurkis<-mean(duomenys$susirgimai); vidurkis 23,0625

Kadangi žmonės neskaičiuojami po kablelio, atsakymą suapvaliname:

round(vidurkis, 0); vidurkis 23

Randame duomenų medianą:

median(duomenys$susirgimai) 12

Randame duomenų standartinį nuokrypį:

sd(duomenys$susirgimai) 25.47192

**3**

Duomenų išrinkimas padarytas darbo pradžioje

**4**

Duomenis skirstome į savaitinius. Pirma randame, kiek iš viso yra duomenų eilučių:

n<-length(duomenys$susirgimai) 80

Tada tikrinsime kiek turėsime savaičių:

nsav<-ceiling(n/7)

Kiekvienai dienai priskiriame po savaitės numerį (kelintai savaitei priklauso)

sav<-rep (1 : nsav, each=7, length.out=n)

Sukuriame lentelę vuzualizuojančią susirgimų skaičių sumą priskiriamą kiekvienai savaitei

tapply (duomenys$susirgimai, sav, sum)

1 2 3 4 5 6 7 8 9 10 11 12

88 62 66 180 99 280 227 359 305 162 16 1