Domača naloga 1

Neza Krzan, Tom Rupnik

Podatki

Uporabila bova podatke *Swiss banknotes data*, ki vsebujejo šest meritev, opravljenih na 100 pravih in 100 ponarejenih starih švicarskih bankovcih za 1000 frankov.

Podatki vsebujejo 7 spremenljivk - 6 številskih in eno opisno. Vsebujejo različne izmerjene dolžine in širine bankovca v milimetrih:

- `length`: dolžina bankovca(na sliki x_1),
- `left`: dolžina levega roba(na sliki x_2),
- `right`: dolžina desnega roba(na sliki x_3),
- `bottom`: dolžina spodnjega roba(na sliki x_4) in
- `top`: dolžina zornjega roba(na sliki x_5) ter
- `diag`: dolžina diagonale bankovca(na sliki x_6).

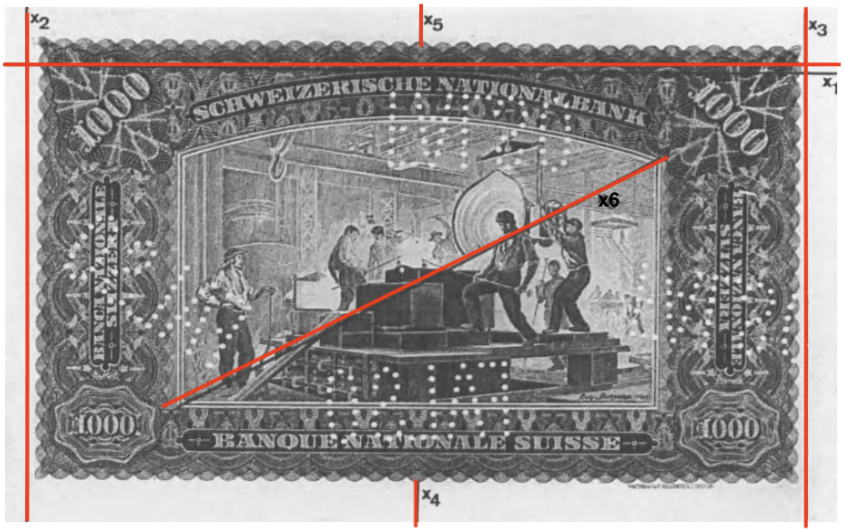


Figure 1: Označene mere na bankovcu.

Opisna spremenlivka `status` pa določa ali je bankovec pravi(`genuine`) ali ponarejen(`counterfeit`). V tabeli imamo torej meritve za 200 različnih bankovcev.

Urejanje podatkov

Imena spremenljivk in vrednosti kategorične spremenljivke sva preimenovala v slovenska imena ter, kot sva že napisala zgoraj, sva podatke skalirala.

Preimenovane spremenljivke:

- `length`: `dolžina`,

Table 1: Opisne statistike za številske spremenljivke v podatkovnem okviru `Swiss banknotes data`.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 50	Pctl. 75	Max
dolzina	200	215	0.4	214	215	215	215	216
levi.rob	200	130	0.4	129	130	130	130	131
desni.rob	200	130	0.4	129	130	130	130	131
spodnji.rob	200	9	1	7	8	9	11	13
zgornji.rob	200	11	0.8	8	10	11	11	12
diagonala	200	140	1	138	140	140	142	142

- `left`: `levi.rob`,
- `right`: `desni.rob`,
- `bottom`: `spodnji.rob`,
- `top`: `zgornji.rob`,
- `diag`: `diagonala` in
- `status` : `tip`, kjer je potem `counterfeit`:`ponarejen bankovec` in `genuine`:`pravi bankovec`.

Za lažjo predstavo si poglejmo opisne statistike številskih spremenljivk, da bomo vedeli s kakšnimi podatki imamo opravka.

Spremenljivke imajo različen razpon vrednosti, zato jih bova, skalirala; vidimo pa tudi, da nimamo mankajočih vrednosti v podatkih.

Poglejmo si še porazdelitve spremenljivk.

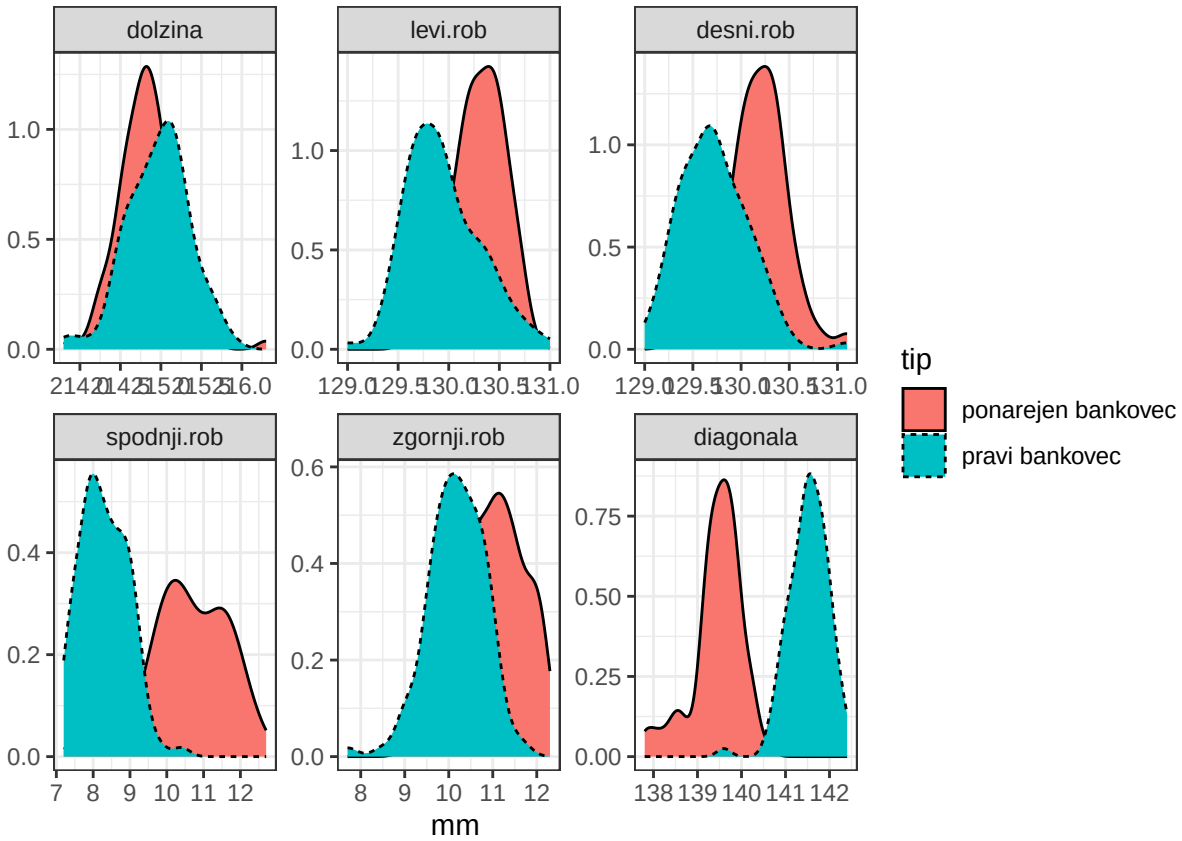


Figure 2: Porazdelitve spremenljivk v podatkovnem okviru `Swiss banknotes data`.

Opazna je razlika med pravimi bankovci in ponarejenimi pri vseh spremenljivkah.

Za razvrščanje bova uporabljala samo številske spremenljivke, in sicer `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob`, `zgornji.rob`; za analizo pa spremenljivki `tip` in `diagonala`. Ker je `diagonala` edina številska spremenljivka pri analizi, le ta ne bo skalirana.

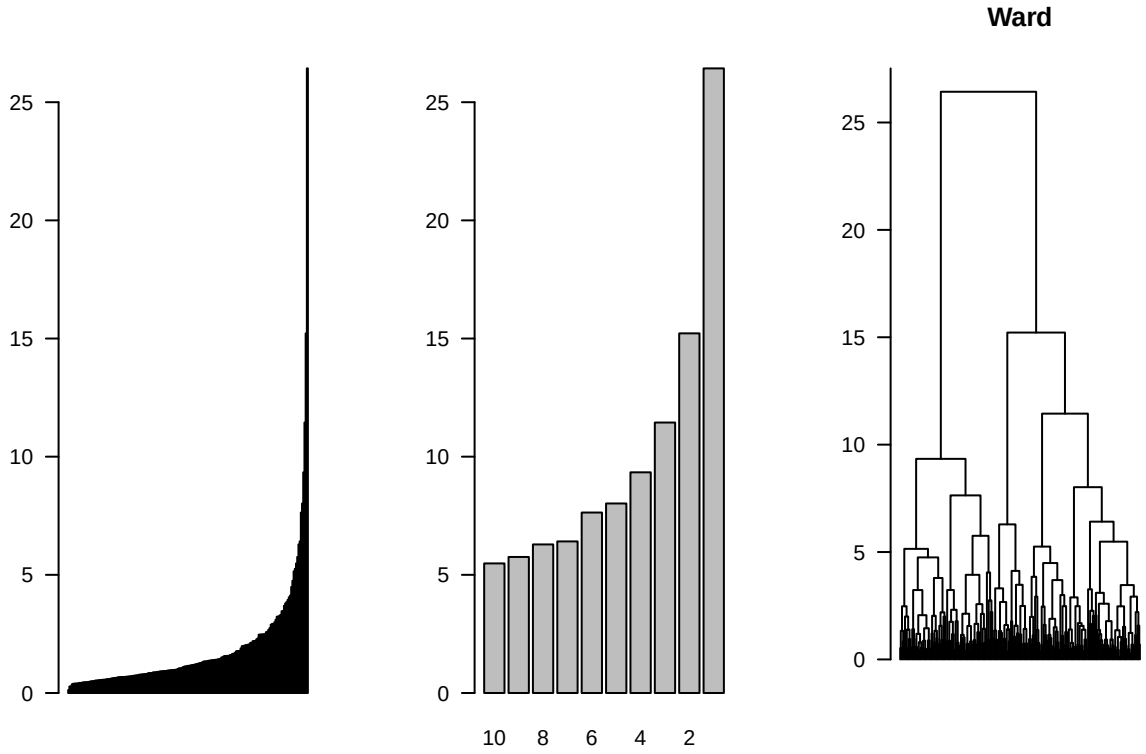
Hierarhično razvrščanje

Torej za razvrščanje uporabljava spremenljivke `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob` in `zgornji.rob` ter primerjala bova tri različne metode in sicer, Wardovo metodo, minimalno metoda (single linkage) in maksimalno metoda (complete linkage).

Wardowa metoda

```
# matrika razliznosti na standardiziranih podatkih (Evklidska razdalja)
dz <- dist(x=dfz, method="euclidean")

hc.ward <- hclust(d=dz, method="ward.D2")
oldpar <- par(las=1, mfrow=c(1, 3))
barplot(hc.ward$height)
barplot(tail(x=hc.ward$height, n=10), names.arg=rev(seq_len(10)))
plot(hc.ward, labels=F, hang=-1, main="Ward", sub="", xlab="", ylab="")
```




```
par(oldpar)
```