Analiza podataka i obrada informacija

Primjer kolokvija #1

Kolokvij nosi ukupno **30 bodova** i piše se **120 minuta**.

Na pitanja odgovarajte tako da pitanja kopirate u kôd i zakomentirate, **ispod svakog napišete odgovarajuće naredbe**, a **ispod naredbi kao komentar upišete odgovore**. Kôd treba pokazati kako ste došli do rezultata.

- **Obavezno navedite redni broj pitanja!**

Priprema podataka (5 bodova)

1. **(1 bod)** Učitajte podatke **mjerenja.xlsx** kao data frame naziva **mjere**. Prema potrebi instalirajte i aktivirajte odgovarajući paket. Podaci se odnose na tjelesne mjere ispitanika i njihove rezultate u dizanju utega.

2. **(1 bod)** Postavite sjeme na vaš **JMBAG** te slučajnim odabirom odaberite **20 redaka** koje ćete izbaciti. Rezultat odabira kopirajte u kôd.

3. **(1 bod)** Slučajnim odabirom odaberite **5 stupaca** koje ćete zadržati. Birajte između:

    - `"Podlaktica", "Biceps", "Prsa", "Vrat", "Ramena", "Struk", "List", "Bedro", "Glava"`
    - Rezultat odabira kopirajte u kôd.

4. **(1 bod)** U odabrane stupce, u podatcima trebate ostaviti i stupce:

    - `"Ime", "Visina", "Tezina", "Dizanje utega1", "Dizanje utega2"`

5. **(1 bod)** Izbacite nepotrebne retke.

Spremite preostale podatke kao CSV datoteku pokretanjem naredbe:

```
write.csv(mjere, file = paste(JMBAG, "_df.csv", sep = ""), row.names = TRUE)
```

Zadatak 1. (10 bodova)

Odgovorite:

6. **(1 bod)** Koliko ima **kategorijskih**, a koliko **numeričkih** varijabli? Navedite nazive svih kategorijskih varijabli.
7. **(1 bod)** Koliko ima **ispitanika**?
8. **(1 bod)** Napravite **tablicu frekvencija imena**. Koja imena se ponavljaju više od jedan put?
9. **(2 bod)** Izračunajte **raspon** vrijednosti numeričkih varijabli. Koja varijabla ima najveći raspon vrijednosti?

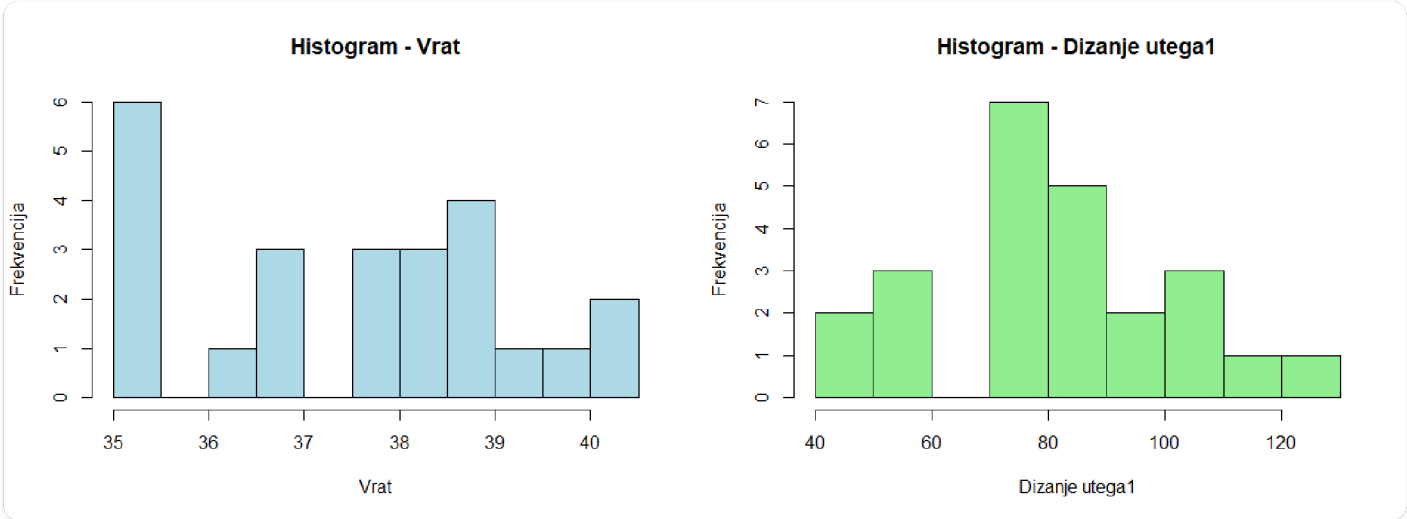
10. **(2 boda)** Izračunajte **BMI** (*body mass index*) svakog ispitanika i rezultat dodajte kao novi stupac naziva **BMI**. Za izračun koristite formulu:

$$\text{BMI} = \frac{\text{težina (kg)}}{(\text{visina (m)})^2}$$

11. **(1 bod)** Koji ispitanici imaju **BMI veći od 27**? Navedite njihova imena.
12. **(2 boda)** Prikažite pomoću **histograma** s 10 razreda distribuciju varijabli s **najmanjim i najvećim rasponom vrijednosti** iz zadatka br. 9.

- Grafičke prikaze postavite **jedan pored drugoga**
- Dodajte **naslove i potrebne oznake**

Primjer:



Zadatak 2. (10 bodova)

13. **(1 bod)** Izračunajte broj **nedostajućih vrijednosti** po svakoj varijabli.
14. **(3 boda)** Ako se u nekoj varijabli nalazi **više od 5% nedostajućih vrijednosti**, zamijenite ih **aritmetičkom sredinom**. Ako ih je manje od **3%**, izbacite ih.
15. **(2 boda)** U novom **okviru** nazvanom **napredni** izdvojite sve ispitanike koji su na **prvom mjerenju** (Dizanje utega 1) mogli dići više od vlastite težine. Preostale ispitanike izdvojite u drugi okvir naziva **potencijalni**.
16. **(4 boda)** Ako u svakoj skupini (**potencijalni** i **napredni**) izdvojite **najlošijeg** ispitanika po **prvom mjerenju** (Dizanje utega 1), te usporedite njihove rezultate na **drugom mjerenju** (Dizanje utega 2), koji je od njih ostvario bolji napredak? Izračunajte i odgovorite.

Zadatak 3. (5 bodova)

17. **(2 boda)** Izračunajte **standardnu devijaciju** za sve numeričke varijable. Koja varijabla ima **najveću standardnu devijaciju**?

18. **(1 bod)** Normalizirajte podatke tako da svaka numerička varijabla ima srednju vrijednost **0** i standardnu devijaciju **1**. Koristite **Z-score normalizaciju**:

$$\mathrm{Z} = \frac{X - \bar{X}}{\sigma}$$

19. **(2 bod)** Izdvojite ispitanike koji imaju **Z-score** podatke za varijable `Visina` i `Tezina` blizu **0**, tj. unutar raspon `[-0.25, 0.25]`. Navedite njihova imena.

Predajte sljedeće datoteke:

- **CSV datoteku** s vlastitim podacima
- `R` ili `Rmd` datoteku
- **Opcionalno**:
  - **JPG datoteku** s grafičkim prikazom
  - **Rezultate izvođenja** u `PDF` formatu