ANOVA & Multiple Testing
Gegeven de volgende tabel van gemiddelde uitgezeten tijd in Amerikaanse gevangenissen voor verschillende misdaden.

Misdaad	**n**	**Gemiddelde x_i**	**s_i**
Geweld	16	14.5	4.5
Drugs	16	18.2	4.2
Vuurwapens	16	18.4	4.5
Fraude	16	11.5	4.7

(0.67 punten) Pas ANOVA toe om te bepalen of we kunnen verwerpen dat alle gemiddelde uitgezeten tijden hetzelfde zijn. Gebruik alpha = 0.05 (tabel met kritieke waarden van de F-verdeling gegeven voor alpha=0.05)

(0.33 punten) Indien we bij de vorige vraag zagen dat de gemiddeldes inderdaad verschillend zijn, pas dan Tukey's procedure toe om te bepalen welke verschillend zijn. (tabel gegeven met kritieke waarden om die Q_{alpha, I, I(J-1)} te berekenen)

Recommendatiesystemen
(0.33 punten) Je wilt een recommendatiesysteem opzetten voor een online boekenwinkel die net gestart is, met 1.000.000 boeken en al 10.000 reviews. Welk van volgende systemen zou hier het beste werken? Motiveer je antwoord

- User-based recommendatie
- Item-based recommendatie
- Content-based recommendation

(0.67 punten) Zijn de volgende beweringen waar of vals? Geen uitleg gevraagd.

- Voor collaboratieve filtering heb je altijd expliciete feedback (bv beoordelingen) van gebruikers nodig.
- Het verhogen van het aantal neighbours bij user-based recommendatiesystemen leidt altijd tot een hogere nauwkeurigheid van de aanbevelingen.
- Een user-based recommendatiesysteem kan een lage precision en een hoge recall hebben.
- Pearson correlation en cosine similarity zullen altijd leiden tot dezelfde rangschikking van neighbours van een item of gebruiker.
- Een oplossing voor het cold-start probleem is het gebruiken van een gemengd recommendatiesysteem.

Clustering
Gegeven een kleine (n=10) tweedimensionale dataset op een scatterplotje.

(0.33 punten) Wat is het verschil tussen hierarchische en k-means clustering?

(0.33 punten) Op bovenstaande dataset zijn drie hierarchische clustermethodes toegepast, namelijk Single linkage, Complete linkage, en Ward linkage, en de drie dendrogrammen zijn gegeven. Gevraagd: welk dendrogram hoort bij welke clustermethode? (+verklaar)

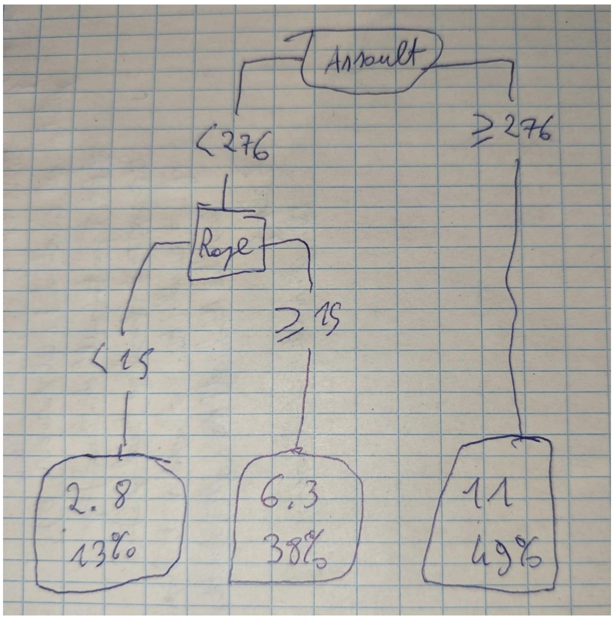
(0.33 punten) Als we het eerste dendrogram gebruiken en het volgende R commando runnen: cutree(hcA, 3). Welke clusters zullen dan vormen? Duid aan op de scatterplot.

Beslissingsbomen
Het doel is om aan de hand van het aantal arrestaties per 100000 voor Assault, Rape en de variabele UrbanPop (% stedelijke bevolking) van een Amerikaanse staat, aan de hand van een beslissingsboom het aantal arrestaties per 100000 voor moord (Murder) te vinden (uit de USArrests dataset uit de werkcolleges).

Gegeven R-commando voor het opstellen van een boom tree1 met minsplit=2, cp=0, en een boom tree2 met cp=cp.min (ook commando voor het bepalen van cp.min natuurlijk) met crossvalidatie, en de berekening van de RMSE's van beide bomen.
Output: RMSE tree1 = 2.9, RMSE tree2 = 2.6

(0.67 punten) Verklaar het verschil in RMSE tussen beide bomen. Leg hierbij ook uit wat de parameters van de R commandos betekenen en hoe crossvalidatie hierbij komt kijken.

(0.33 punten) Gegeven een Rpart.plot van tree2, leg uit hoe we aan de hand hiervan voorspelingen kunnen maken voor Murder aan de hand van Assault, Rape en UrbanPop.



Recreatie van de gegeven figuur, cijfers zijn mss niet 100% accuraat.

Permutatie en rang testen
Dataset: dataset met gegevens over wagens
Doel: bepalen of er een verschil is in gemiddelde PK tussen auto's met manuele schakeling en auto's met automatische schakeling.
Gegeven:

- Boxplot van PK's van auto's met manuele schakeling en auto's met automatische schakeling, hierop was te zien dat manuele schakeling grote outliers had maar dat de medianen wel ver van elkaar lagen.
- Het feit dat er 19 auto's waren met automatische en 13 met manuele schakeling.
- Code en output voor de Wilk-Shapiro normaliteitstest op PK's van automatische auto's (p=0.44)
- Code en output voor de Wilk-Shapiro normaliteitstest op PK's van manuele auto's (p=0.003)
- Code en output voor de Levene test op de varianties van manuele en automatische auto's (p=0.2)
- Code en output voor de t-test, een (zelf geschreven) permutatietest (p=0.167), en de Wilcoxon rang sum test (p=0.0457)

(0.33 punten) Wat proberen de t-test, permutatietest en Wilcoxon test te testen? Geef de nulhypothese en de alternatieve hypothese. Kunnen we op basis van deze gegevens hier een t-test doen of is het aangewezen een andere test te gebruiken?

(0.33 punten) Leg aan de hand van de gegeven code uit wat de permutatietest doet. Wat kun je concluderen uit het resultaat van deze test?

(0.33 punten) Wat kan je concluderen uit de Wilcoxon rank sum test? Verklaar het verschil tussen dit resultaat en het resultaat van de permutatietest.

Principale Componenten Analyse
Gegeven: R-code en output voor het toepassen van PCA op de 'bev' dataset, een dataset met gegevens over elektrische wagens. Specifiek de volgende dingen:

- Gemiddelde en SD van elke kolom uit de dataset
- Eigenwaarden bij PCA op covariantiematrix van de dataset
- Eigenwaarden bij PCA op correlatiematrix van de dataset
- Biplot van de PCA op de correlatiematrix

(0.33 punten) Is het hier handiger om PCA toe te passen op de covariantiematrix of op de correlatiematrix? Verklaar.

(0.33 punten) Hoeveel procent van de variantie wordt verklaard door de eerste 2 principale componenten?

(0.33 punten) Interpreteer de eerste 2 principale componenten aan de hand van de gegeven biplot.