

Examen Statistiek II: Project

M. Luyts & A. Huybrechts

Academiejaar 2024-2025

Naam + Voornaam:

Studentennummer:

Instructies voor het examen:

- U krijgt 3 uur voor het examen, inclusief het mondelinge examen dat maximaal 10 min duurt.
- Het examen is open boek.
- U mag gebruik maken van uw eigen notities, slides van de cursus en documentatie rond de gegeven projecten. GSM's en laptops zijn niet toegelaten.
- Rekenmachines zijn toegelaten.
- Vragen die we stellen gaan over de topics die we behandeld hebben tijdens de lessen.
- Gelieve uw naam zowel op deze examenkopij te schrijven (op elk blad), als op uw oplossingenbladen (op elk blad). **Het is verplicht om alles in te dienen: ook deze examenkopij, uw kladpapier en alle notities die u tijdens het examen maakt.**
- Beantwoord elke vraag op een duidelijke, gestructureerde manier. Wees beknopt, precies en to-the-point. Het is niet altijd nodig om volledige zinnen te schrijven, lijsten met opsommingstekens kunnen voldoende zijn.
- Puntenverdeling: Schriftelijk examen (6 pt) en mondeling examen (2 pt op de mondelinge verdediging van Project 3)
- Spieken en/of fraude is ten striktste verboden!

Alvast veel succes!

Handtekening:

.....

Vraag 1. [ANOVA en Multiple testing; 15 punten]

Het Amerikaanse Bureau van gevangenen publiceert gegevens over de verblijfsduur van gevangenen voor het eerst vrijgelaten uit federale instellingen. Onafhankelijke willekeurige metingen van vrijgegeven gedetineerden voor vier verschillende delictcategorieën leverden de volgende informatie op over de uitgezeten tijd, in maanden.

Categorie	n_i	$\bar{y}_i$	s_i
Namaak	16	14.5	4.5
Drugs	16	18.4	3.8
Vuurwapens	16	18.2	4.5
Fraude	16	11.5	4.7

- [10 punten] Is er op basis van de variantieanalyse een verschil in het gemiddelde uitgezeten tijd tussen de vier delictgroepen? Gebruik een significantieniveau van 5%. Hieronder vindt u de kritische waarden terug van een $F_{DF1,DF2}$ -verdeling voor $\alpha = 0.05$.

F-table of Critical Values of $\alpha = 0.05$ for F(df1, df2)																					
	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞		
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31		
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53		
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63		
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37		
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23		
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71		
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54		
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40		
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30		
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21		
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13		
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07		
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01		
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96		
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92		
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88		
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84		
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81		
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78		
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76		
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73		
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71		
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69		
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67		
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65		
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64		
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62		
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51		
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39		
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25		
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00		

2. [5 punten] Pas de Tukey-test procedure toe om de gemiddelde verschillen te vergelijken als de ANOVA-test uit 1 significant is op het 5%-niveau. Welke conclusies kan je hieruit trekken?

De onderstaande tabel geeft de kritische waarden weer voor de studentized range $q(\alpha = 0.05, k, DF)$, met k het aantal groepen en $DF = (k.n) - k$ met n het totaal aantal waarnemingen per groep.

Critical Values of Studentized Range Distribution(q) for Familywise ALPHA = .05.

Denominator DF	Number of Groups (a.k.a. Treatments)							
	3	4	5	6	7	8	9	10
1	26.976	32.819	37.081	40.407	43.118	45.397	47.356	49.070
2	8.331	9.798	10.881	11.734	12.434	13.027	13.538	13.987
3	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462
4	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826
5	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995
6	4.339	4.896	5.305	5.629	5.895	6.122	6.319	6.493
7	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158
8	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918
9	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738
10	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598
11	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486
12	3.773	4.199	4.508	4.748	4.947	5.116	5.262	5.395
13	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318
14	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253
15	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150
17	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071
19	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037
20	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008
21	3.565	3.942	4.213	4.424	4.597	4.743	4.870	4.981
22	3.553	3.927	4.196	4.405	4.577	4.722	4.847	4.957
23	3.542	3.914	4.180	4.388	4.558	4.702	4.826	4.935
24	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
25	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897
26	3.514	3.880	4.141	4.345	4.511	4.652	4.773	4.880
27	3.506	3.870	4.130	4.333	4.498	4.638	4.758	4.864
28	3.499	3.861	4.120	4.322	4.486	4.625	4.745	4.850
29	3.493	3.853	4.111	4.311	4.475	4.613	4.732	4.837
30	3.487	3.845	4.102	4.301	4.464	4.601	4.720	4.824
31	3.481	3.838	4.094	4.292	4.454	4.591	4.709	4.813
32	3.475	3.832	4.086	4.284	4.445	4.581	4.698	4.802
33	3.470	3.825	4.079	4.276	4.436	4.572	4.689	4.791
34	3.465	3.820	4.072	4.268	4.428	4.563	4.680	4.782
35	3.461	3.814	4.066	4.261	4.421	4.555	4.671	4.773
36	3.457	3.809	4.060	4.255	4.414	4.547	4.663	4.764
37	3.453	3.804	4.054	4.249	4.407	4.540	4.655	4.756
38	3.449	3.799	4.049	4.243	4.400	4.533	4.648	4.749
39	3.445	3.795	4.044	4.237	4.394	4.527	4.641	4.741
40	3.442	3.791	4.039	4.232	4.388	4.521	4.634	4.735
41	3.439	3.787	4.035	4.227	4.383	4.515	4.628	4.728
42	3.436	3.783	4.030	4.222	4.378	4.509	4.622	4.722
43	3.433	3.779	4.026	4.217	4.373	4.504	4.617	4.716
44	3.430	3.776	4.022	4.213	4.368	4.499	4.611	4.710
45	3.428	3.773	4.018	4.209	4.364	4.494	4.606	4.705
46	3.425	3.770	4.015	4.205	4.359	4.489	4.601	4.700
47	3.423	3.767	4.011	4.201	4.355	4.485	4.597	4.695
48	3.420	3.764	4.008	4.197	4.351	4.481	4.592	4.690
49	3.418	3.761	4.005	4.194	4.347	4.477	4.588	4.686
50	3.416	3.758	4.002	4.190	4.344	4.473	4.584	4.681

Critical Values of Studentized Range Distribution(q) for Familywise ALPHA = .05.

Denominator DF	Number of Groups (a.k.a. Treatments)							
	3	4	5	6	7	8	9	10
51	3.414	3.756	3.999	4.187	4.340	4.469	4.580	4.677
52	3.412	3.753	3.996	4.184	4.337	4.465	4.576	4.673
53	3.410	3.751	3.994	4.181	4.334	4.462	4.572	4.669
54	3.408	3.749	3.991	4.178	4.331	4.459	4.569	4.666
55	3.406	3.747	3.989	4.176	4.328	4.455	4.566	4.662
56	3.405	3.745	3.986	4.173	4.325	4.452	4.562	4.659
57	3.403	3.743	3.984	4.170	4.322	4.449	4.559	4.656
58	3.402	3.741	3.982	4.168	4.319	4.447	4.556	4.652
59	3.400	3.739	3.979	4.165	4.317	4.444	4.553	4.649
60	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
61	3.397	3.735	3.975	4.161	4.312	4.439	4.548	4.643
62	3.396	3.734	3.973	4.159	4.309	4.436	4.545	4.641
63	3.395	3.732	3.972	4.157	4.307	4.434	4.542	4.638
64	3.393	3.730	3.970	4.155	4.305	4.431	4.540	4.635
65	3.392	3.729	3.968	4.153	4.303	4.429	4.538	4.633
66	3.391	3.727	3.966	4.151	4.301	4.427	4.535	4.630
67	3.390	3.726	3.965	4.149	4.299	4.425	4.533	4.628
68	3.389	3.725	3.963	4.147	4.297	4.423	4.531	4.626
69	3.387	3.723	3.962	4.146	4.295	4.421	4.529	4.624
70	3.386	3.722	3.960	4.144	4.293	4.419	4.527	4.621
71	3.385	3.721	3.959	4.142	4.291	4.417	4.525	4.619
72	3.384	3.719	3.957	4.141	4.290	4.415	4.523	4.617
73	3.383	3.718	3.956	4.139	4.288	4.413	4.521	4.615
74	3.382	3.717	3.954	4.138	4.286	4.411	4.519	4.613
75	3.382	3.716	3.953	4.136	4.285	4.410	4.517	4.611
76	3.381	3.715	3.952	4.135	4.283	4.408	4.515	4.610
77	3.380	3.714	3.951	4.133	4.282	4.406	4.514	4.608
78	3.379	3.713	3.949	4.132	4.280	4.405	4.512	4.606
79	3.378	3.712	3.948	4.131	4.279	4.403	4.511	4.604
80	3.377	3.711	3.947	4.129	4.278	4.402	4.509	4.603
81	3.377	3.710	3.946	4.128	4.276	4.400	4.507	4.601
82	3.376	3.709	3.945	4.127	4.275	4.399	4.506	4.600
83	3.375	3.708	3.944	4.126	4.274	4.398	4.504	4.598
84	3.374	3.707	3.943	4.125	4.272	4.396	4.503	4.597
85	3.374	3.706	3.942	4.123	4.271	4.395	4.502	4.595
86	3.373	3.705	3.941	4.122	4.270	4.394	4.500	4.594
87	3.372	3.704	3.940	4.121	4.269	4.392	4.499	4.592
88	3.372	3.704	3.939	4.120	4.268	4.391	4.498	4.591
89	3.371	3.703	3.938	4.119	4.266	4.390	4.496	4.590
90	3.370	3.702	3.937	4.118	4.265	4.389	4.495	4.588
91	3.370	3.701	3.936	4.117	4.264	4.388	4.494	4.587
92	3.369	3.700	3.935	4.116	4.263	4.387	4.493	4.586
93	3.368	3.700	3.934	4.115	4.262	4.386	4.492	4.585
94	3.368	3.699	3.934	4.114	4.261	4.384	4.491	4.583
95	3.367	3.698	3.933	4.114	4.260	4.383	4.489	4.582
96	3.367	3.698	3.932	4.113	4.259	4.382	4.488	4.581
97	3.366	3.697	3.931	4.112	4.258	4.381	4.487	4.580
98	3.366	3.696	3.930	4.111	4.257	4.380	4.486	4.579
99	3.365	3.696	3.930	4.110	4.257	4.379	4.485	4.578
100	3.365	3.695	3.929	4.109	4.256	4.379	4.484	4.577

Vraag 2. [Recommendatie systemen; 15 punten]

1. [5 punten] U wilt een recommendatie systeem ontwerpen voor een online boekwinkel dat onlangs gelanceerd werd. De boekwinkel heeft meer dan 1 miljoen boektitels, maar de beoordelingsdatabase heeft slechts 10000 beoordelingen. Welke van de volgende mogelijkheden zou een beter recommendatie systeem zijn?

- (a) User-user collaborative filtering
- (b) Item-item collaborative filtering
- (c) Content-based recommendatie

Onderbouw uw antwoord met een duidelijke, korte motivering.

2. [10 punten] Welke uitspraken zijn waar/niet waar?

- (a) Voor collaborative filtering recommendatie systemen is expliciete feedback (bijvoorbeeld beoordelingen) van gebruikers op items vereist.
- (b) Een grotere neighbourhood op user-based collaboratieve filtering recommendatie systemen impliceert altijd een betere nauwkeurigheid van de aanbevelingen.
- (c) Een collaboratieve filtering kan een hoge recall hebben, terwijl de precision laag is.
- (d) Het gebruik van Pearson-correlatie en cosine similarity leidt altijd tot dezelfde rangschikking van vergelijkbare gebruikers in collaboratieve filtering.
- (e) Het cold-start probleem kan worden aangepakt door verschillende recommendatie systemen te combineren.

Vraag 3. [Principale componenten analyse; 15 punten]

Men heeft een PCA uitgevoerd op de dataset “bev” over elektrische auto’s, die de volgende gegevens bevat: Prijs (in euro), Acceleratie 0 - 100 km/u (in sec), Topsnelheid (in km/u), Actieradius Elektrisch (in km), Totaal Vermogen (in kW), Lengte (in mm), Breedte (in mm), Hoogte (in mm), Bagageruimte (in L) en Massa Rijklaar (in kg). Op basis van deze data voeren de onderzoekers een eerste verkennende analyse uit en vervolgens een principale componenten analyse.

```
> apply(bev, 2, mean)
```

Prijs (in euro)	Acceleratie 0 - 100 km/u (in sec)
67799.15625	7.88625
Topsnelheid (in km/u)	Actieradius Elektrisch (in km)
177.52500	345.18750
Totaal Vermogen (in kW)	Laadsnelheid (in km/u)
212.25625	46.41875
Lengte (in mm)	Breedte (in mm)
4736.05625	1885.94375
Hoogte (in mm)	Bagageruimte (in L)
1629.55000	532.11250
Massa Rijklaar (in kg)	
2098.64375	

```
> apply(bev, 2, sd)
```

Prijs (in euro)	Acceleratie 0 - 100 km/u (in sec)
35627.943260	3.132734
Topsnelheid (in km/u)	Actieradius Elektrisch (in km)
39.387440	99.954215
Totaal Vermogen (in kW)	Laadsnelheid (in km/u)
125.183752	8.005086
Lengte (in mm)	Breedte (in mm)
362.536734	65.183268
Hoogte (in mm)	Bagageruimte (in L)
162.735994	213.971108
Massa Rijklaar (in kg)	
304.475926	

```
> eigen(cov(bev))$values
```

```
[1] 1.269451e+09 1.654641e+05 2.880837e+04 1.569904e+04 1.127738e+04 3.254804e+03  
[7] 1.591547e+03 8.106139e+02 1.042346e+02 1.708714e+01 6.165362e-01
```

```
> eigen(cov(bev))$vectors[,1:2]
```

```
      [,1]      [,2]  
[1,]  9.999602e-01 -0.005491027  
[2,] -5.104752e-05  0.003783450
```

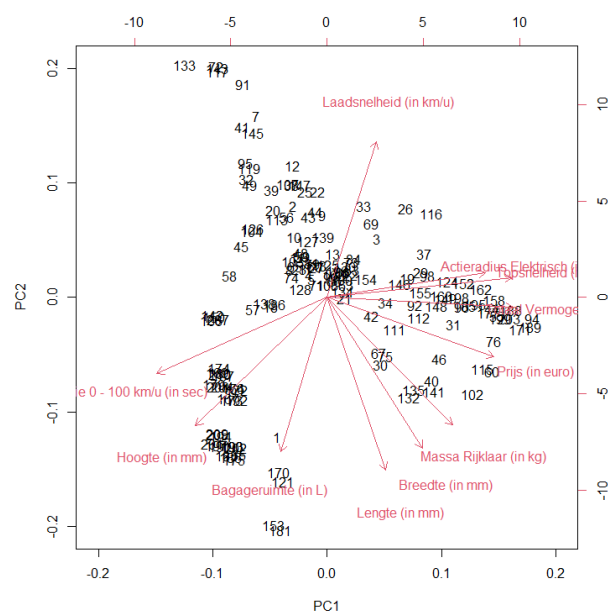
```
[3,] 8.422639e-04 -0.025050278
[4,] 1.608690e-03 -0.032922373
[5,] 3.024527e-03 -0.053711996
[6,] -1.151866e-05 -0.012931143
[7,] 5.164865e-03 0.738876531
[8,] 1.176617e-03 0.092091875
[9,] -1.698334e-03 0.278434914
[10,] -1.866288e-04 0.456855058
[11,] 6.005935e-03 0.393102984
```

```
> eigen(cor(bev))$values
```

```
[1] 5.29739700 3.77722146 0.66442772 0.36501194 0.29200060 0.19755629 0.17038942
[9] 0.09117071 0.06334621 0.05439066 0.02708798
```

```
> eigen(cor(bev))$vectors[,1:2]
```

```
      [,1]      [,2]
[1,] 0.3726564 -0.15615412
[2,] -0.3797666 -0.19979398
[3,] 0.4131327 0.05253834
[4,] 0.3536920 0.06514034
[5,] 0.4144903 -0.02537329
[6,] 0.1098756 0.41174110
[7,] 0.1307623 -0.45519764
[8,] 0.2128362 -0.39688897
[9,] -0.2932717 -0.33832668
[10,] -0.1024308 -0.40770633
[11,] 0.2799189 -0.33631093
```



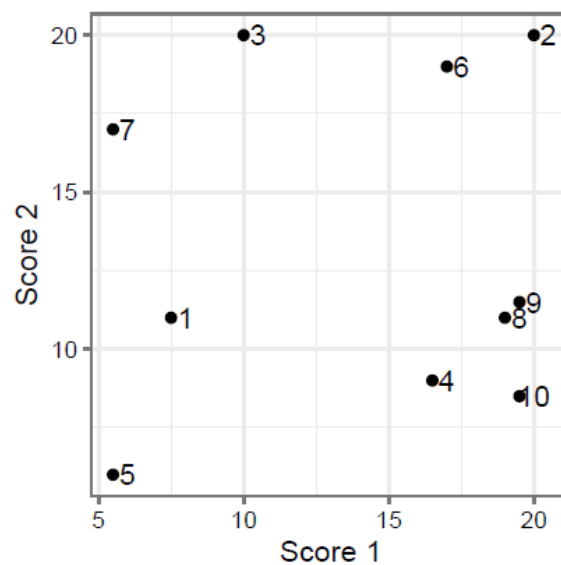
Beantwoord de volgende vragen:

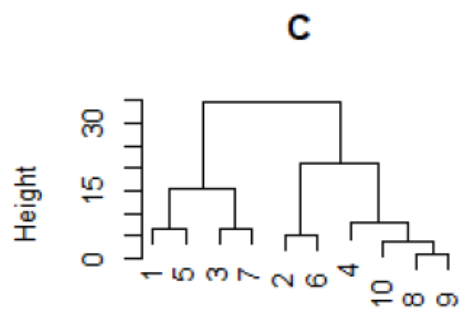
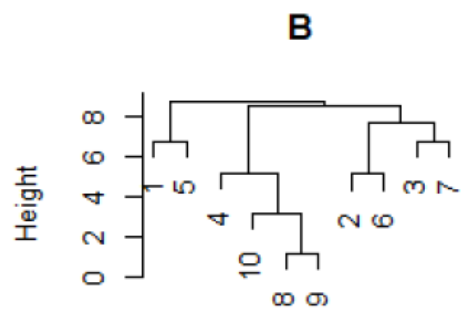
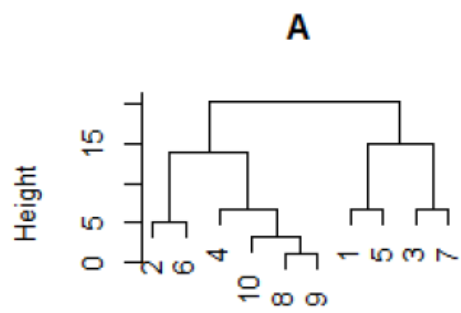
1. [5 punten] In een PCA analyse voor deze data, zou het beter zijn om de covariantie- of de correlatiematrix te gebruiken? Motiveer jouw antwoord.
2. [5 punten] Welk percentage van de totale variantie bevatten de eerste twee hoofdcomponenten?
3. [5 punten] Interpreteer de eerste twee hoofdcomponenten aan de hand van de biplot.

Vraag 4. [Clustering; 15 punten]

1. [5 punten] Wat is het verschil tussen hierarchische clustering en k-means clustering?
2. [5 punten] Gegeven de scores voor 2 examens van 10 leerlingen. Deze worden weergegeven op onderstaande figuur. Deze leerlingen werden met hierarchische clustering geclusterd met 3 verschillende methodes: single linkage, complete linkage en ward linkage. Welke clustering methode werd in welke figuur gebruikt? Leg kort uit waarom.
3. [5 punten] Welke clusters zouden in A gevormd worden na gebruik van deze code? Duid aan op de scatterplot.

```
cluster <- cutree(hcA, 3)
```





Vraag 5. [Beslissingsbomen & Cross-validatie; 15 punten]

Bij deze vraag zullen we gebruik maken van de dataset 'USArrests'. Deze bevat data over arrestaties voor overvallen, moorden en verkrachtingen per 100000 inwoners in de 50 Amerikaanse staten in 1973. Daarnaast bevat deze dataset ook informatie over het percentage van de populatie dat in stedelijke gebieden woont.

- Murder: aantal arrestaties voor moord (per 100000)
- Assault: aantal arrestaties voor overvallen (per 100000)
- Rape: aantal arrestaties voor verkrachting (per 100000)
- UrbanPop: percentage stedelijke populatie

```
data("USArrests")
head(USArrests)
```

##	Murder	Assault	UrbanPop	Rape
## Alabama	13.2	236	58	21.2
## Alaska	10.0	263	48	44.5
## Arizona	8.1	294	80	31.0
## Arkansas	8.8	190	50	19.5
## California	9.0	276	91	40.6
## Colorado	7.9	204	78	38.7

Bij deze zijn we geïnteresseerd in het voorspellen van het aantal arrestaties voor moord, gegeven de cijfers voor overval, verkrachting en het percentage van de populatie dat in stedelijke gebieden leeft. Hiervoor wordt een kansboom opgesteld. Dit stukje code splitst de dataset eerst op in een training dataset en een test dataset.

```
set.seed(2001)
ind<-sample(50,40)
training<-USArrests[ind, ]
test<-USArrests[-ind,]
```

Met het volgende stukje code worden 2 beslissingsbomen gemaakt. De resulterende RMSE voor elke boom wordt weergegeven als output.

```
library(rpart)
library(rpart.plot)
tree1<-rpart(Murder~., data=training, control=rpart.control(minsplit=2,cp=0))
cp.min <- tree1$cptable[which.min(tree1$cptable[, "xerror"]), "CP"]
tree2 <- prune(tree1, cp=cp.min)
```

```

pred1<-predict(object=tree1, newdata=test)
rmse1<-sqrt(mean((pred1-test$Murder)^2))
pred2<-predict(object=tree2, newdata=test)
rmse2<-sqrt(mean((pred2-test$Murder)^2))

data.frame(rmse1,rmse2)

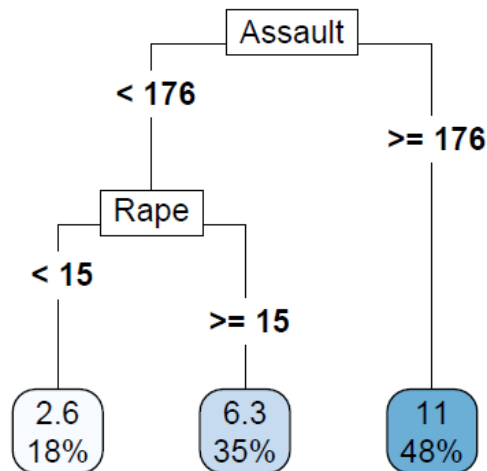
```

```

##      rmse1    rmse2
## 1 2.895514 2.671753

```

1. [10 punten] Verklaar het verschil in RMSE. Leg hierbij ook uit welke parameters in het opstellen van de boom werden gebruikt. Wat is het belang van cross-validatie?
2. [5 punten] Deze plot geeft de resulterende boom weer. Leg uit hoe je deze beslissingsboom kan gebruiken om een voorspelling te doen over het aantal arrestaties voor 'Murder', gegeven 'Assault', 'Rape' en 'UrbanPop'.



Vraag 6. [Permutatie- en rangtesten; 15 punten]

Voor deze vraag zijn we geïnteresseerd in de 'mtcars' dataset. Deze dataset bevat gegevens over 32 auto's. We willen onderzoeken of er een verschil is in 'horsepower' (hp) tussen auto's met automatische versus manuele schakeling (am; 0 = automatisch, 1 = manueel).

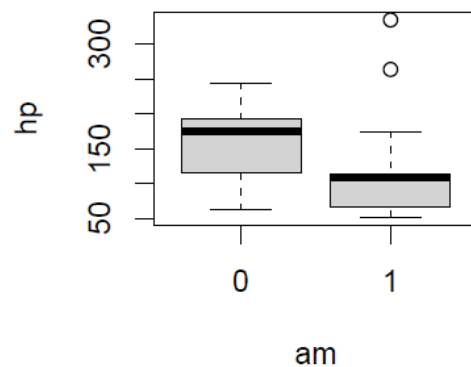
```
data(mtcars)
head(mtcars)
```

```
##           mpg cyl  disp  hp  drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive  21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant         18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
attach(mtcars)
table(am)
```

```
## am
##  0  1
## 19 13
```

```
boxplot(hp~am)
```



Shapiro-Wilk test en levene test:

```
shapiro.test(hp[am==0])
```

```
##  
## Shapiro-Wilk normality test
```

```
##  
## data: hp[am == 0]  
## W = 0.95835, p-value = 0.5403
```

```
shapiro.test(hp[am==1])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: hp[am == 1]  
## W = 0.76758, p-value = 0.00288
```

```
car::leveneTest(hp~as.factor(am))
```

```
## Levene's Test for Homogeneity of Variance (center = median)  
##      Df F value Pr(>F)  
## group 1  0.1842 0.6709  
##      30
```

T-test:

```
t.test(hp~am, var.equal=T)
```

```
##  
## Two Sample t-test  
##  
## data: hp by am  
## t = 1.3733, df = 30, p-value = 0.1798  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -16.27768 83.11169  
## sample estimates:  
## mean in group 0 mean in group 1  
##      160.2632      126.8462
```

Permutatietest:


```

T<-mean(hp[am==0])-mean(hp[am==1])
S<-10000
set.seed(2001)
teststatistieken<-c()
data <- mtcars
for(i in 1:S){
  data$am <- sample(data$am)
  teststatistieken[i]<-mean(data$hp[data$am==0])-
    mean(data$hp[data$am==1])
}
pwaarde<-mean(abs(teststatistieken)>abs(T))
T

```

```
## [1] 33.417
```

```
pwaarde
```

```
## [1] 0.174
```

Wilcoxon rank sum test:

```
wilcox.test(hp~am)
```

```

##
## Wilcoxon rank sum test with continuity correction
##
## data: hp by am
## W = 176, p-value = 0.0457
## alternative hypothesis: true location shift is not equal to 0

```

1. [5 punten] Welke hypothese wordt getest in de 3 gegeven testen? Geef zowel de nulhypothese als de alternatieve hypothese. Op basis van de gegevens, volstaat het om een t-test uit te voeren voor het testen van deze nulhypothese, of kiezen we beter een alternatief zoals de permutatietest en de Wilcoxon test? Verklaar.
2. [5 punten] Leg uit met behulp van de gegeven code hoe de permutatietest werkt. Wat is de conclusie van deze permutatietest?
2. [5 punten] Wat is de conclusie van de Wilcoxon test? Vergelijk met de permutatietest (en verklaar).

