Universität Regensburg

P.D. Dr. David Elsweiler

WHK "Einführung in die Informationswissenschaft" SoSe 2019

Nachname, Vorname	
Abschluss (BA, MA, FKN etc.)	
Matrikelnummer, Semester	
Versuch (1/2/3)	

Bitte füllen Sie zuerst den Kopf des Angabenblattes aus!

In diesem Klausurteil sind insgesamt 30 Punkte zu erreichen. Der Klausurteil besteht aus fünf Seiten.

Es sind keine Hilfsmittel zugelassen.

Die Klausur besteht aus 5 Aufgaben.

Bitte beantworten Sie alle Fragen direkt auf das Angabenblatt.
Nutzen Sie ggf. die Rückseite und kennzeichnen Sie dies entsprechend.
Eigene Schmierblätter sind nicht erlaubt.

Bei mehreren oder mehrdeutigen Lösungen wird die schlechtere Lösung gewertet. Streichen Sie daher ungültige Lösungen eindeutig durch.

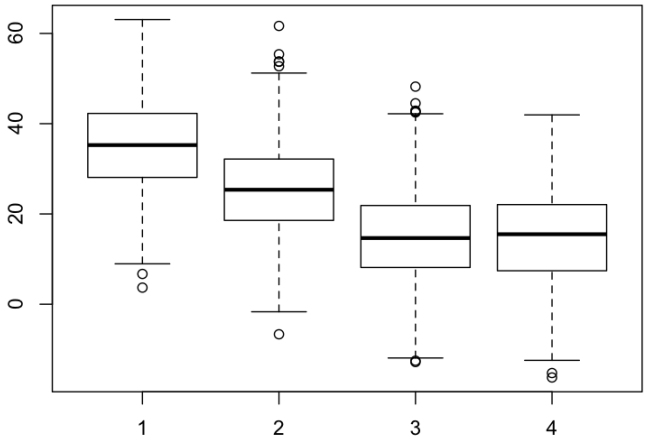
Viel Erfolg!

Aufgabe 1: (3 Punkte)
Calculate the mode, mean, and median for the following set of numbers. Show your working, *Berechnen Sie den Modus, den arithmetischen Mittelwert und den Median für die unten gelisteten Zahlen*

9	8	8	7	5	7	8	8	5	12

Mode	
Mean	
Median	

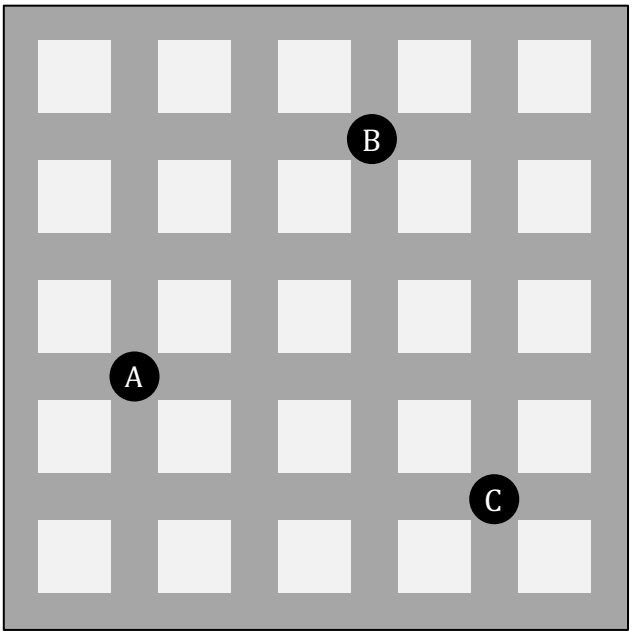
Aufgabe 2: (5 Punkte)
Four distributions are shown in the Figure below. We perform t-tests to compare distributions 1 and 2 (i.e. test 1 = t.test(dist1, dist2)) and 3 and 4 (i.e. test2 = t.test (dist3,dist4)). *Die Grafik unten zeigt 4 Verteilungen. Wir benutzen t-tests um Dist 1 mit Dist 2 und Dist 3 mit Dist 4 zu vergleichen.*



Which of the following statements are true: Incorrect answers will result in minus points. Minimum point score for the question is 0.
Welche der folgenden Aussagen treffen zu? Die Angabe widersprüchlicher Lösungen führt zur Bewertung mit 0 Punkten.

Zutreffend		**nicht zutreffend**
☐	The t-value for test 1 > t-value for test 2 *der t-Wert für Test 1 > der t-Wert für Test 2*	☐
☐	The p-value for test 1 > p-value for test 2 *der p-Wert für Test 1 > der p-Wert für Test 2*	☐
☐	There are no outliers in distribution 4 *Es gibt keine Ausreißer in Verteilung 4*	☐
☐	Test 2 is more likely to produce a significant result than test 1 *Es ist wahrscheinlicher, dass Test 2 ein signifikantes Ergebnis liefern würde*	☐
☐	All four distributions are distributed Gaussian (normal) *alle vier Verteilungen sind normal-verteilt.*	☐

Aufgabe 3: (7 Punkte)



a) What is the Manhattan distance between A and B ?
Wie groß ist die Manhattan Abstand zwischen A und B?

2 2.8 3 4 5

b) What is the Manhattan distance between B and C ?
Wie groß ist die Manhattan Abstand zwischen B und C?

3.1 4 4.5 6 10

c) What is the Euclidean distance between B and C?
Was ist der euklidische Abstand zwischen B und C?

1 2.8 3.2 4.3 5

d) What is the Euclidean distance between A and B?
Was ist der euklidische Abstand zwischen A und B?

1 2 3.1 4.3 5

e) Are linear regression, hierarchical clustering and k-means all examples of unsupervised learning approaches?
Sind lineare Regression, hierarchische Clusteranalyse und k-means alle unüberwachte Lernverfahren?

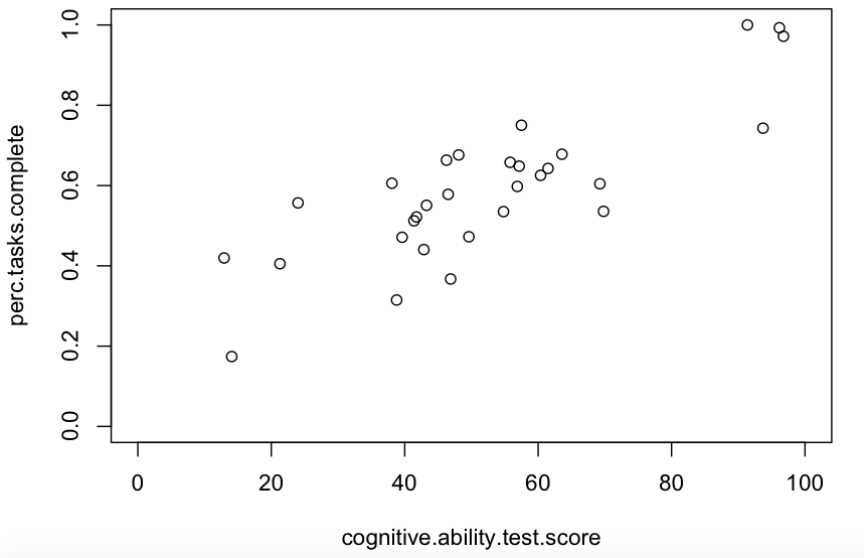
Yes No

f) Name 2 parameters required to run k-means clustering
Nennen Sie 2 Parameter, die nötig sind um eine k-means-Analyse durchzuführen

g) Will repeating k-means clustering with the same parameters always produce the same clusters?
Liefert eine k-means-Analyse mit gleichen Input-Parametern immer dasselbe Ergebnis?

Yes No

Aufgabe 4: (8 Punkte)



a) Fit an estimated regression model to the data plotted above e.g. $h_\theta(x) = \theta_0 + \theta_1 x_1$ by drawing it on the diagram.
Schätzen Sie ein Regressionsmodell und zeichnen Sie es auf die oben gezeigte Grafik

Based on your model answer the following questions:
Beantworten Sie die unten gelisteten Fragen basierend auf Ihrem Modell

b) Do the data suggest that the more tasks a participant completed the higher they scored on the cognitive ability test?
Deuten die Daten an, dass das Ergebnis beim Test der kognitiven Leistung eines Teilnehmers umso höher ist je mehr Aufgaben er/sie geschafft hat?

Yes No

c) Which of the following is true? *Welche der nächsten Aussagen stimmt?*
$\theta_1 > 0$,
$\theta_1 < 0$
$\theta_1 == 0$,

d) How many of the tasks would someone with the mean test score be expected to complete?
Wie viele der Aufgaben würden Sie gemäß dem Modell erwarten, dass ein Proband mit dem durchschnittlichen Testergebnis erledigen würde?

10- 40%
40- 60%
60-100%

Aufgabe 5: (3+4 Punkte)

A retrieval system returns 15 documents for a query it estimates to be relevant out of a collection of 10,000. 5 of these are actually relevant. 45 relevant documents exist. A user reads half of the returned documents. Estimate Precision and Recall for this example.

Ein IR-System schätzt von einer Sammlung mit 10,000 Dokumente 15 als relevant ein. 5 der zurückgelieferten Dokumente sind tatsächlich relevant. Es gibt 45 relevante Dokumente in der Sammlung. Der Nutzer liest nur die Hälfte der als relevant geschätzten Dokumente. Berechnen Sie Precision und Recall.

b) Create an inverted-index for the following documents.
Erstellen Sie einen „inverted-index" für die unten gelisteten Dokumente.

D1: We were having fun at the party

D2: The conservatives are the nasty party

D3: The little boy was very nasty to his sister even at her birthday party