

Bachelorprüfung Data Science: Datenauswertung (21791), Wintersemester 2020/21

Liebe Studierende,

markieren Sie bitte Ihre Antworten auf dem Antwortbogen am Ende des Gehefts in der folgenden Weise: .

Wenn Sie eine Antwort korrigieren möchten, füllen Sie bitte die **falsch** markierte Antwort vollständig aus, ungefähr so: .

Bitte füllen Sie folgende Angaben deutlich lesbar aus:

Nachname : _____

Vorname : _____

Matrikelnummer : _____

Studiengang : _____

Raum, Platz : _____

Prüfer : Prof. Dovern

WICHTIG: Bitte kreuzen Sie Ihre Matrikelnummer auch auf dem Antwortbogen an!

Nachfolgende Angaben sind nur vom Prüfer auszufüllen:

Note:

Unterschrift Prüfer:

Bitte beachten Sie folgende Hinweise:

- Das Geheft **muss** zusammen bleiben!
- Die Klausur besteht aus insgesamt 25 **Single-Choice-Fragen**, von denen 6 R-Bezug haben.
- Verwenden Sie für Ihre Antworten ausschließlich den Antwortbogen am Ende des Gehefts.
Einträge in der Aufgabenstellung werden nicht gewertet!
- Beschriften Sie den Antwortbogen deutlich lesbar mit Ihrem Namen und Ihrer Matrikelnummer und kreuzen Sie Ihre Matrikelnummer dort zusätzlich an!
- Verwenden Sie auf dem Antwortbogen bitte einen **dunklen Kugelschreiber!**
- Bearbeitungszeit: 60 Minuten
- **Erlaubte Hilfsmittel:**
 - Nicht-programmierbarer Taschenrechner
 - Die vom Lehrstuhl offiziell herausgegebene Formelsammlung, 2. bis 4. Auflage, ohne weitere Eintragungen oder Markierungen, mit Ausnahme von farblichen Hinterlegungen von Textpassagen und/oder Formeln bzw. unbeschriebenen Post-Its
 - Cheat Sheet für Basics in R, das über StudOn bereitgestellt wurde, ohne weitere Eintragungen oder Markierungen, mit Ausnahme von farblichen Hinterlegungen von Textpassagen und/oder Befehlen

Viel Erfolg!

MUSTER
Nicht ausfüllen!

Bachelorprüfung Data Science: Datenauswertung, WiSe 2020/21

Aufgabe 1

Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.

Hinweis: Aufgabe 1 besteht aus 19 Teilaufgaben, bei denen jeweils ein Punkt erreicht werden kann. Jede Frage bietet mehrere Antwortmöglichkeiten, von denen **jeweils nur eine korrekt ist**. Kreuzen Sie jeweils die korrekte Antwort **auf dem Antwortbogen** an. Beachten Sie, dass es **keinen Punktabzug für falsch beantwortete Fragen** gibt.

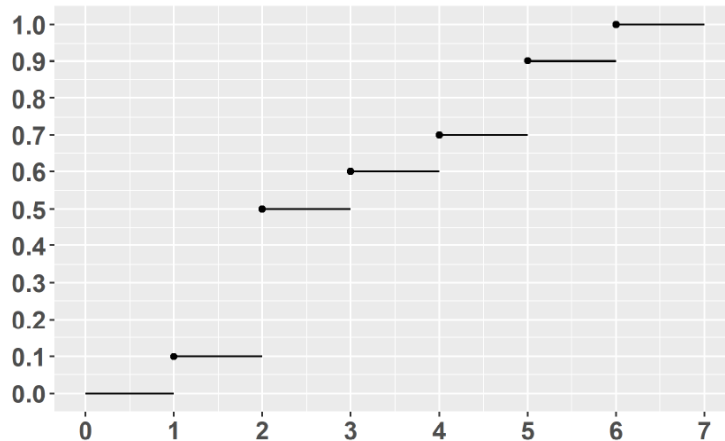
- 1.1** Ordnen Sie den folgenden drei Merkmalen der Reihenfolge nach ihre Skalenniveaus zu:
Anzahl an abgeschlossenen Fachsemestern, Schulabschluss (z.B. Hauptschule, Realschule, Abitur), Intelligenzquotient
- A** Absolutskala, Ordinalskala, Intervallskala
 - B** Verhältnisskala, Ordinalskala, Verhältnisskala
 - C** Absolutskala, Nominalskala, Intervallskala
 - D** Verhältnisskala, Nominalskala, Absolutskala
 - E** Absolutskala, Absolutskala, Absolutskala

Betrachten Sie die folgende Kontingenztafel für die beiden Merkmale X und Y basierend auf einer Erhebungsgesamtheit der Größe 100:

| | | Y | | | $\sum_{j=1}^3 y_j$ |
|--------------------|-------|-------|-------|-------|--------------------|
| | | y_1 | y_2 | y_3 | |
| X | x_1 | 15 | 10 | 5 | 30 |
| | x_2 | 5 | ? | 5 | ? |
| | x_3 | 10 | 10 | 20 | 40 |
| $\sum_{i=1}^3 x_i$ | | ? | ? | 30 | ? |

- 1.2** Wie lauten die fehlenden gemeinsamen Häufigkeiten und Randhäufigkeiten?
- A** $n_{22} = 20, n_{2\cdot} = 30, n_{\cdot 1} = 30, n_{\cdot 2} = 40$
 - B** $n_{22} = 20, n_{\cdot 2} = 30, n_{\cdot 1} = 30, n_{2\cdot} = 40$
 - C** $n_{22} = 30, n_{2\cdot} = 40, n_{\cdot 1} = 50, n_{\cdot 2} = 30$
 - D** $n_{22} = 30, n_{\cdot 2} = 40, n_{\cdot 1} = 50, n_{2\cdot} = 30$
 - E** Die fehlenden Werte lassen sich nicht ohne zusätzliche Informationen bestimmen.

Betrachten Sie die folgende empirische Verteilungsfunktion, die auf einer sehr großen Erhebungsgesamtheit beruht:



- 1.3 Welche Aussage über die dargestellte Verteilung ist **nicht** korrekt?
- A Der Modus der Verteilung hat den Wert 2.
 - B Der Wert des Medians ist kleiner als der Wert des Modus.
 - C Das zugrundeliegende Merkmal folgt einer diskreten Verteilung.
 - D Vier Merkmalsausprägungen haben die gleiche relative Häufigkeit.
 - E Vier Merkmalsausprägungen haben die gleiche absolute Häufigkeit.

10000 Autobesitzer wurden nach der Fahrleistung ihres PKW im Jahr 2019 (in 1000 km) gefragt. Dabei ergab sich die folgende Häufigkeitsverteilung:

| Klasse | Gefahrene km in 1000 | Anzahl der Autobesitzer | h_i | \hat{F}_i |
|----------------|----------------------|-------------------------|-------|-------------|
| i | $[x_{i-1}^*, x_i^*)$ | n_i | | |
| 1 | $[5, 20)$ | 6000 | 0.6 | 0.6 |
| 2 | $[20, 50)$ | 3000 | 0.3 | 0.9 |
| 3 | $[50, 100)$ | 1000 | 0.1 | 1.0 |
| $\sum_{i=1}^3$ | | 10000 | 1.0 | |

- 1.4 Wie hoch ist der Anteil der Autobesitzer die im Jahr 2019 maximal 30000 km gefahren sind, wenn Sie von einer Gleichverteilung innerhalb der Klassen ausgehen?
- A 30%
 - B 50%
 - C 70%
 - D 80%
 - E Der gesuchte Wert lässt sich mit den gegebenen Informationen nicht berechnen.

1.5 Gegeben sei eine Erhebung vom Umfang n für die beiden Merkmale X und Y . Für jede Beobachtung x_i und y_i , $i = 1, \dots, n$, gelte $z_i = x_i + y_i$. Darüber hinaus bezeichnen \bar{x} , \bar{y} bzw. \bar{z} die dazugehörigen arithmetischen Mittel. $\sum_{i=1}^n (x_i - \bar{x})$ bezeichnet die Summe der mittelwertbereinigten Daten. Welche der folgenden Aussagen ist korrekt?

- A** $i) \bar{z} = \bar{x} + \bar{y}$ und $ii) \sum_{i=1}^n (x_i - \bar{x}) = 0$
B $i) \bar{z} \neq \bar{x} + \bar{y}$ und $ii) \sum_{i=1}^n (x_i - \bar{x}) \neq 0$
C $i) \bar{z} = \bar{x} + \bar{y}$ und $ii) \sum_{i=1}^n (x_i - \bar{x}) \neq 0$
D $i) \bar{z} \neq \bar{x} + \bar{y}$ und $ii) \sum_{i=1}^n (x_i - \bar{x}) = 0$
E $i) \bar{z} = \bar{x} + \bar{y}$ und $ii) \sum_{i=1}^n (x_i - \bar{x})^2 = 0$

1.6 Gegeben seien Stichproben x_i und $y_i = a + bx_i$, $i = 1, \dots, n$, wobei a und b reelle Zahlen sind. Welche der folgenden Aussagen ist korrekt?

- A** $s_y^2 = a + bs_x^2$
B $s_y^2 = bs_x^2$
C $s_y = a + bs_x$
D $s_y^2 = b^2 s_x^2$
E $s_x^2 = b^2 s_y^2$

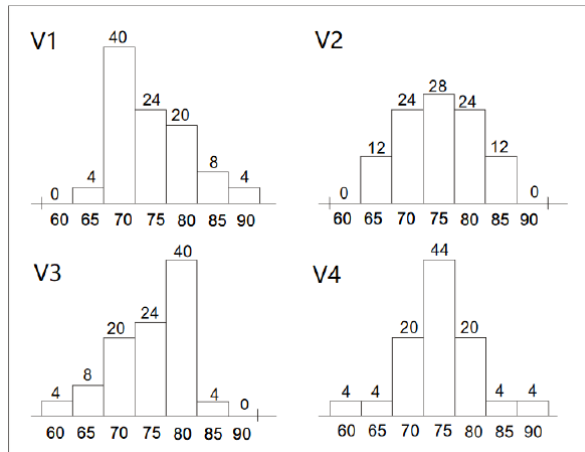
Bei einer Erhebung wurden 16 Studierende, $i = 1, \dots, 16$, gefragt, wie viele Tassen Kaffee sie pro Tag trinken (Merkmal K). Die Ergebnisse liegen in der folgenden Urliste vor:

| | | | | | | | | | | | | | | | | |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|
| i : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| k_i : | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 5 | 5 |

1.7 Welche Werte nehmen die Spannweite s_M und der Interquartilsabstand s_Q an?

- A** $s_M = 3$ und $s_Q = 5$
B $s_M = 5$ und $s_Q = 4$
C $s_M = 4$ und $s_Q = 3$
D $s_M = 5$ und $s_Q = 3$
E $s_M = 3$ und $s_Q = 4$

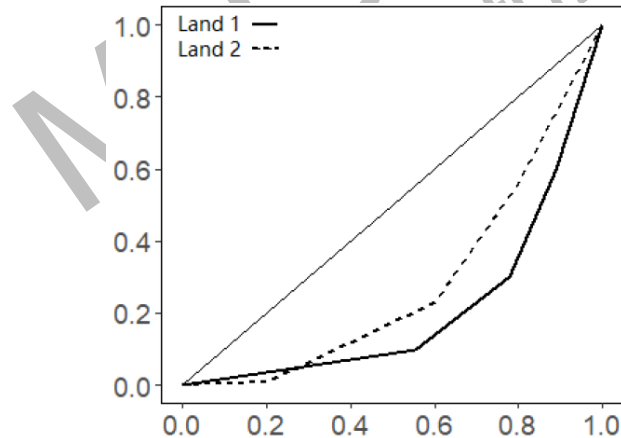
Die folgenden vier Häufigkeitsverteilungen (V1, V2, V3, V4) haben jeweils das gleiche arithmetische Mittel und die gleiche Varianz. Gleichwohl ist ihre Gestalt hinsichtlich Schiefe und Wölbung sehr unterschiedlich.



1.8 Welche der folgenden Aussagen ist richtig?

- A V2 und V4 haben die gleiche Kurtosis
- B Schiefe von V2 < Schiefe von V3
- C Schiefe von V3 = (-1) · Schiefe von V1
- D Kurtosis von V1 > Kurtosis von V3
- E Die Schiefe von V2 und V3 ist gleich

Das nachfolgende Schaubild zeigt die Lorenzkurven der Vermögensverteilung in zwei Ländern.



1.9 Im Folgenden sei G_i der Gini-Koeffizient für Land $i = 1, 2$. Welche Aussage über die dargestellte Vermögensverteilung ist korrekt?

- A Die Konzentrationsfläche von Land 2 ist größer als die von Land 1.
- B $G_1 < G_2$
- C $G_1 = G_2$
- D Die Vermögensungleichheit in Land 1 ist geringer als jene in Land 2.
- E $G_1 > G_2$

- 1.10** Auf einem Markt sind drei Unternehmen tätig. Zwei davon haben einen Marktanteil von je 25%. Zur Konzentrationsmessung soll der Herfindahl-Index, H , bestimmt werden. Welchen Wert nimmt H im vorliegenden Beispiel an?
- A** $H = 0.500$
B $H = 0.333$
C $H = 0.375$
D $H = 0.187$
E $H = 0.125$

Eine zufällige Stichprobe von $n = 100$ Unternehmensberatern/-beraterinnen umfasst studierte Ökonomen/Ökonominnen (Ö), Naturwissenschaftler/-innen (N) und Geisteswissenschaftler/-innen (G). Alle wurden nach ihrem Einkommen befragt. Dann wurde bestimmt, ob das Einkommen einer Person über dem Durchschnittseinkommen liegt (\bar{U}), oder nicht (U). In der folgenden Kontingenztabelle fehlt eine Häufigkeit:

| | Ökonom (Ö) | Naturw. (N) | Geistw. (G) |
|--------------------|------------|-------------|-------------|
| Über (\bar{U}) | 26 | ? | 11 |
| Unter (U) | 24 | 13 | 9 |

- 1.11** Der Anteil derjenigen, die ein Studium der Naturwissenschaften abgeschlossen haben und überdurchschnittlich verdienen, ist gegeben durch:

- A** $h(\bar{U}, N) = 0.170$
B $h(N|\bar{U}) = 0.314$
C $h(N, \bar{U}) = 0.130$
D $h(\bar{U}|N) = 0.170$
E $h(\bar{U}|N) = 0.566$

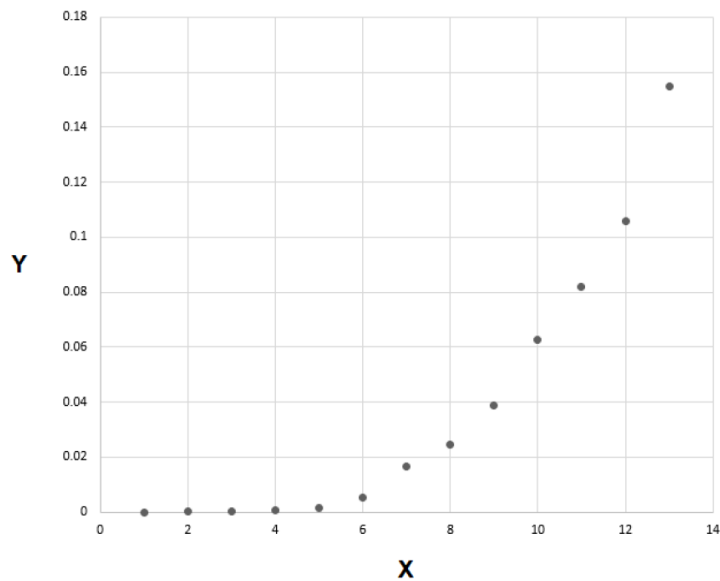
- 1.12** Der Anteil der Unternehmensberater/-innen, die überdurchschnittlich verdienen, unter denjenigen, die ein Studium entweder der Naturwissenschaften oder der Geisteswissenschaften abgeschlossen haben, ist gegeben durch:

- A** $\frac{h(N)+h(G)}{h(\bar{U})}$
B $h(\bar{U}, G) + h(\bar{U}, N)$
C $\frac{h(\bar{U}, G)+h(\bar{U}, N)}{h(G)+h(N)}$
D $h(\bar{U}|G) + h(\bar{U}|N)$
E $\frac{h(\bar{U}|G)+h(\bar{U}|N)}{h(G)+h(N)}$

- 1.13** Welche der folgenden Aussagen zum Korrelationskoeffizient nach Pearson ist allgemein korrekt?

- A** $r_{XY} = r_{YX}$
B $r_{XY} = 1 - r_{YX}$
C $r_{XY} = -r_{YX}$
D $0 \leq r_{XY} \leq 1$
E Keine der anderen Aussagen ist korrekt.

Die folgende Graphik zeigt ein Streudiagramm für zwei Merkmale X und Y :



1.14 Welche der folgenden Aussagen ist korrekt?

- A Der Korrelationskoeffizient nach Pearson ist geeignet, um den dargestellten Zusammenhang sinnvoll zu erfassen.
- B Der Korrelationskoeffizient nach Pearson ist im vorliegenden Beispiel negativ.
- C Der Korrelationskoeffizient nach Pearson ergibt im vorliegenden Beispiel eine höhere Korrelation als der Korrelationskoeffizient nach Spearman.
- D Die Korrelation nach Spearman hat für den dargestellten Zusammenhang den Wert 0.
- E Keine der anderen Antworten ist korrekt.

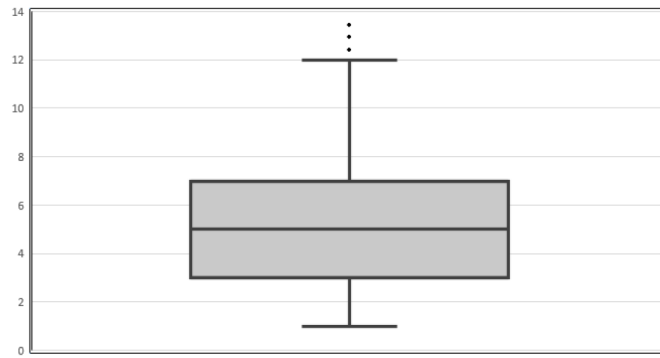
Gegeben sind die folgenden Wettkampfergebnisse für fünf Sportler:

| i | 1 | 2 | 3 | 4 | 5 |
|-------------|----|----|----|----|----|
| Platzierung | 1 | 3 | 5 | 4 | 2 |
| Alter | 18 | 21 | 24 | 25 | 22 |

1.15 Berechnen Sie den Zusammenhang zwischen dem verhältnisskalierten Alter und der ordinalskalierten Platzierung in einem Wettkampf mit einem geeigneten Maß. Welches der folgenden Ergebnisse ist richtig?

- A 0.80
- B 0.87
- C 0.55
- D 0.67
- E 0.41

Beispielhaft ist im Folgenden ein Boxplot zu sehen:



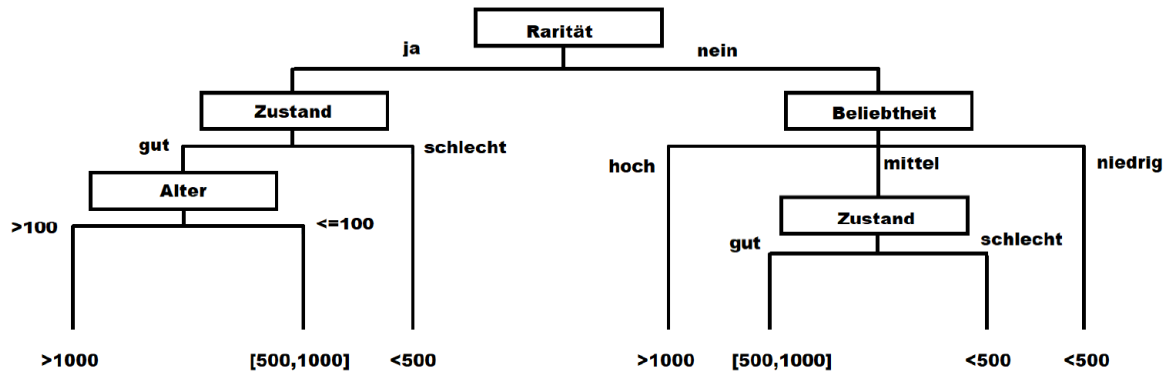
1.16 Welche der folgenden allgemeinen Aussagen über Boxplots ist falsch?

- A** Beobachtungen außerhalb der Antennen bzw. Fühler ("Whisker") werden als Ausreißer bezeichnet.
- B** Die Position des Striches innerhalb in der Box repräsentiert üblicherweise das arithmetische Mittel.
- C** Die Box liegt nicht immer symmetrisch um den Strich innerhalb der Box.
- D** Anhand eines Boxplots lassen sich Ausreißer identifizieren.
- E** Die Größe der Box wird durch die Quartile bestimmt.

1.17 Sie haben einen Datensatz mit Informationen zu 300 Features für $n = 20000$ Social-Media-Profile und möchten diese in möglichst homogene Gruppen einteilen. Welche Methode eignet sich dafür?

- A** Neuronales Netz
- B** Random Forest
- C** K-means-Clustering
- D** Principal-Component-Analyse
- E** Support Vector Machine

Der Besitzer eines Antiquariats möchte seine Ankäufe optimieren. Dafür möchte er die Preise für die Antiquitäten voraussagen, wobei er diese in drei Kategorien einteilt (unter 500 Euro, zwischen 500 und 1000 Euro, mehr als 1000 Euro). Zur Vorhersage verwendet er die Merkmale "Rarität", "Zustand", "Alter" (in Jahren) und "Beliebtheit". Betrachten Sie den folgenden Entscheidungsbaum, der zur Preisprognose herangezogen werden soll:



1.18 Welche Aussage ist falsch?

- A Für eine 80 Jahre alte rare Antiquität in gutem Zustand wird ein Verkaufspreis zwischen 500 und 1000 Euro prognostiziert.
- B Für eine nicht-rare Antiquität von mittlerer Beliebtheit in schlechtem Zustand wird ein Verkaufspreis von weniger als 500 Euro prognostiziert.
- C Für eine nicht-rare Antiquität von hoher Beliebtheit wird ein Verkaufspreis zwischen 500 und 1000 Euro prognostiziert.
- D Für eine rare Antiquität in einem schlechten Zustand wird ein Verkaufspreis von weniger als 500 Euro prognostiziert.
- E Für eine 200 Jahre alte, rare Antiquität in gutem Zustand wird ein Verkaufspreis von mehr als 1000 Euro prognostiziert.

MUSTER

Nicht ausfüllen!

Gehen Sie von 5 Beobachtungseinheiten aus, deren Merkmalsausprägungen für zwei Merkmale X und Y in der folgenden Tabelle gegeben sind:

| | | | | | |
|-------|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 | 5 |
| x_i | 6 | 5 | 3 | 1 | 9 |
| y_i | 1 | 4 | 4 | 8 | 3 |

1.19 Betrachten Sie die beiden Clusterschwerpunkte $C^A = (2, 6)$ und $C^B = (5, 4)$. Die Distanz einer Beobachtung zu einem beliebigen Schwerpunkt $C = (c_X, c_Y)$ ist gegeben durch die euklidische Distanz:

$$d_i = \sqrt{(x_i - c_X)^2 + (y_i - c_Y)^2},$$

wobei c_X und c_Y die x- bzw. y-Koordinaten eines Schwerpunktes bezeichnen. Im Rahmen einer K-means-Clustering-Analyse sollen die Clusterzugehörigkeiten der einzelnen Beobachtungen bestimmt werden. Welche der folgenden Aussagen ist korrekt?

- A Die Beobachtung $i = 3$ gehört zu Cluster A.
- B Die Beobachtung $i = 3$ gehört zu Cluster B.
- C Die Beobachtung $i = 2$ gehört zu Cluster A.
- D Die Beobachtung $i = 4$ gehört zu Cluster B.
- E Keine der anderen Aussagen ist korrekt.

Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.

MUSTER
Nicht ausfüllen!

Aufgabe 2

Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.

Hinweis: Aufgabe 2 besteht aus 6 Teilaufgaben, bei denen jeweils ein Punkt erreicht werden kann. Jede Frage bietet mehrere Antwortmöglichkeiten, von denen **jeweils nur eine korrekt ist**. Kreuzen Sie jeweils die korrekte Antwort **auf dem Antwortbogen** an. Beachten Sie, dass es **keinen Punktabzug für falsch beantwortete Fragen** gibt.

2.1 Betrachten Sie die folgende Ablaufstruktur:

```
for(i in 1:5){  
  if(i<=2|i>4){  
    print(i^2+2)  
  } else{  
    print(2*i-1)  
  }  
}
```

Welche Werte gibt diese Befehlssequenz in R aus?

- A 3, 6, 5, 7, 27
- B 1, 2, 3, 4, 5
- C 3, 6, 5, 18, 27
- D 3, 5, 6, 7, 27
- E 0, 1, 1, 2, 3

MUSTER
Nicht ausfüllen!

Gehen Sie für die nächsten Fragen von dem folgenden Workspace in R aus. Der Dataframe `df` enthält Informationen von $n = 191$ Ländern aus dem Jahr 2015 zu den folgenden Merkmalen:

Spalte 1: Der Name des Landes (Land)

Spalte 2: Die CO_2 -Emissionen pro Einwohner in Tonnen pro Jahr (`pkco2`)

Spalte 3: Eine Indikatorvariable, die den Wert 1 annimmt für Länder mit überdurchschnittlichen Pro-Kopf-Emissionen, und 0 sonst (`pkco2_hoch`)

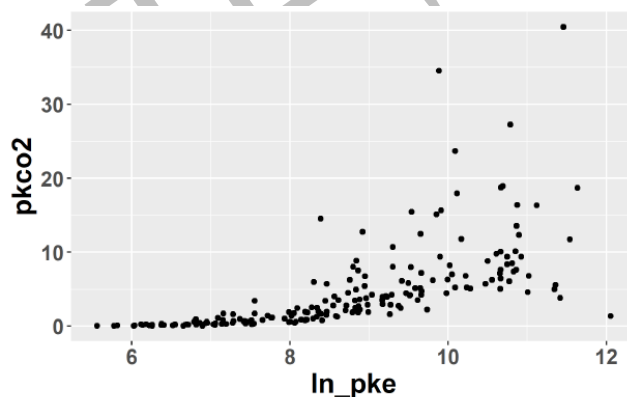
Spalte 4: Das Pro-Kopf-Einkommen in US-Dollar, d.h. das Bruttoinlandsprodukt geteilt durch die Bevölkerungsgröße (`pke`)

Spalte 5: Eine Indikatorvariable, die den Wert 1 annimmt für Länder mit überdurchschnittlichem Pro-Kopf-Einkommen, und 0 sonst (`pke_hoch`)

Spalte 6: Der natürliche Logarithmus des Pro-Kopf-Einkommens (`ln_pke`)

Für jedes Land liegen vollständige Informationen zu allen Merkmalen vor (d.h. es gibt keine NAs). Es gibt keine weiteren Spalten im Dataframe und Sie haben auch sonst keine Datenobjekte (z.B. Values oder Funktionen) abgespeichert. Sie haben das Paket *tidyverse* in Ihrer aktuellen Session bereits aktiviert. Im Rahmen Ihrer Analyse sind die folgenden Tabellen und Schaubilder entstanden:

| | df\$pkco2_hoch | | |
|--------------|----------------|----|-----|
| df\$pke_hoch | 0 | 1 | sum |
| 0 | 119 | 22 | 141 |
| 1 | 5 | 45 | 50 |
| sum | 124 | 67 | 191 |



2.2 Welcher Befehl zeigt Ihnen **nicht** den Anteil der Länder mit überdurchschnittlichen Pro-Kopf-Emissionen an?

- A `mean(df$pkco2_hoch)`
- B `table(df$pkco2_hoch)/length(df$pkco2_hoch)`
- C `sum(df$pkco2_hoch)/length(df$pkco2_hoch)`
- D `summary(df$pkco2_hoch)`
- E `unique(df$pkco2_hoch)`

2.3 Betrachten Sie die oben angezeigte Kontingenztabelle für die beiden Merkmale pke_hoch und pkco2_hoch. Diese wurde mit dem folgenden R-Befehl erzeugt:

```
addmargins(table(df$pke_hoch, df$pkco2_hoch, dnn = c("df$pke_hoch", "df$pkco2_hoch")))
```

Welche der folgenden Aussagen ist **nicht** korrekt?

- A** Die individuellen Beiträge zur Kovarianz zwischen den Merkmalen pke und pkco2 sind mehrheitlich positiv.
- B** Die individuellen Beiträge zur Kovarianz zwischen den Merkmalen pke und pkco2 sind mehrheitlich negativ.
- C** 90% der Länder mit überdurchschnittlichem Pro-Kopf-Einkommen weisen überdurchschnittliche Pro-Kopf-Emissionen auf.
- D** 62.30% der Länder weisen unterdurchschnittliche Pro-Kopf-Einkommen und unterdurchschnittliche Pro-Kopf-Emissionen auf.
- E** 124 Länder im Datensatz weisen unterdurchschnittliche Pro-Kopf-Emissionen auf und 67 Länder haben überdurchschnittliche Pro-Kopf-Emissionen.

2.4 Betrachten Sie die folgende Befehlssequenz:

```
df %>%
  filter(pke>=20000) %>%
  summarize( sqrt( (length(pke)-1)/length(pke) * var(pke) ) )
```

Welches Ergebnis liefert Ihnen diese Befehlssequenz?

- A** Die deskriptive Standardabweichung der Pro-Kopf-Einkommen der Länder mit einem Pro-Kopf-Einkommen von mindestens 20000 US-Dollar.
- B** Die deskriptive Varianz der Pro-Kopf-Einkommen der Länder mit einem Pro-Kopf-Einkommen von mindestens 20000 US-Dollar.
- C** Die deskriptive Standardabweichung der Pro-Kopf-Einkommen aller Länder.
- D** Die deskriptive Varianz der Pro-Kopf-Einkommen aller Länder.
- E** Die deskriptive Standardabweichung der Pro-Kopf-Einkommen der Länder mit einem Pro-Kopf-Einkommen von höchstens 20000 US-Dollar.

2.5 Betrachten Sie das oben angezeigte Streudiagramm für die logarithmierten Pro-Kopf-Einkommen und die Pro-Kopf-Emissionen. Vervollständigen Sie den Befehl

```
ggplot(data = T, aes(U, V)) + geom_W()
```

so, dass das oben angezeigte Streudiagramm erstellt wird.

- A** T: df, U: ln_pke, V: pkco2, W: point
- B** T: df, U: pke, V: pkco2, W: point
- C** T: df, U: pkco2, V: ln_pke, W: point
- D** T: df, U: ln_pke, V: pkco2, W: bar
- E** T: ln_pke, U: pkco2, V: df, W: scatter

2.6 Betrachten Sie die folgende Befehlssequenz:

```
df %>%  
  select(ln_pke, pkco2) %>%  
  kmeans(centers = 4, nstart = 100)
```

Welches Ergebnis liefert Ihnen diese Befehlssequenz?

- A** Die Zugehörigkeit jedes Landes zu einem von vier Clustern auf Basis des K-means-Clustering angewandt auf die Merkmale ln_pke und pkco2.
- B** Die Zugehörigkeit jedes Landes zu einem von 100 Clustern auf Basis des K-means-Clustering angewandt auf die Merkmale ln_pke und pkco2.
- C** Die Zugehörigkeit jedes Landes zu einem von vier Clustern auf Basis des K-means-Clustering angewandt auf die Merkmale pke und pkco2.
- D** Die Zugehörigkeit jedes Landes zu einem von 100 Clustern auf Basis des K-means-Clustering angewandt auf die Merkmale pke und pkco2.
- E** Den Pearson-Korrelationskoeffizient zwischen den Merkmalen ln_pke und pkco2.

Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.

MUSTER
Nicht ausfüllen!

Musterlösung

Bachelorprüfung Data Science:
Datenauswertung, WiSe 2020/21

| | | | | | |
|------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
| 1.1 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.2 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.3 | <input type="checkbox"/> A | <input checked="" type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.4 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.5 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.6 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input checked="" type="checkbox"/> D | <input type="checkbox"/> E |
| 1.7 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input checked="" type="checkbox"/> D | <input type="checkbox"/> E |
| 1.8 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.9 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input checked="" type="checkbox"/> E |
| 1.10 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.11 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.12 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.13 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.14 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input checked="" type="checkbox"/> E |
| 1.15 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.16 | <input type="checkbox"/> A | <input checked="" type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.17 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.18 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input checked="" type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 1.19 | <input type="checkbox"/> A | <input checked="" type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 2.1 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 2.2 | <input type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input checked="" type="checkbox"/> E |
| 2.3 | <input type="checkbox"/> A | <input checked="" type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 2.4 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 2.5 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |
| 2.6 | <input checked="" type="checkbox"/> A | <input type="checkbox"/> B | <input type="checkbox"/> C | <input type="checkbox"/> D | <input type="checkbox"/> E |

MUSTER
Nicht ausfüllen!