

In der Hausübung soll der Datensatz „student_performance.csv“ von der Homepage verwendet werden. In den Daten wurde die Leistung von Schülern an zwei weiterführenden portugiesischen Schulen im Fach Mathematik im Schuljahr 2005/2006 erfasst. Das Schuljahr gliedert sich in drei Trimester. Zum Ende eines jeden Trimesters liegt jeweils eine Note zwischen 0 (=am schlechtesten) und 20 (=am besten) vor.

Der Datensatz enthält die folgenden Variablen:

- school:** student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) - besuchte Schule
- sex:** student's sex (binary: 'F' - female or 'M' - male) - Geschlecht
- age:** student's age (numeric: from 15 to 22) - Alter
- reason:** reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') - Grund, warum diese Schule besucht wird
- G1:** first period grade (numeric: from 0 to 20) - Note im 1. Trimester
- G2:** second period grade (numeric: from 0 to 20) - Note im 2. Trimester
- G3:** final grade (numeric: from 0 to 20) - Note im 3. Trimester (Abschlussnote für das Schuljahr)

Zusatzinformation zu den Daten:

Die Daten sind ein Ausschnitt der Daten, die in folgendem Paper verwendet wurden:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUTURE BUSINESS TECHNOLOGY Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Die Originaldaten sind hier verlinkt und beschrieben:

<http://archive.ics.uci.edu/ml/datasets/Student+Performance>

Die Originaldaten werden aber für die Hausaufgabe nicht benötigt. Bitte verwenden Sie den auf der Homepage verlinkten Datensatz zur Bearbeitung der Hausaufgabe.

Vorbereitung:

- Legen Sie auf dem Z-Laufwerk (im Cip-Pool) einen Ordner für die Hausübung an. Legen Sie alle Dateien, mit denen Sie arbeiten, in diesem Ordner ab. (Hinweis: Die Laufwerke im Cip-Pool werden (voraussichtlich in den Weihnachtsferien) an das LRZ umgezogen. Sie müssen Ihre Daten selbst auf die neuen Laufwerke beim LRZ übertragen. Speichern Sie vorsichtshalber immer alle Ihre Daten auch auf einem USB-Stick.)
- Auf der Homepage ist die Datei „Einf_Software_WS1617_Homepage.R“ verlinkt. Speichern Sie diese in Ihrem Ordner.
- Versuchen Sie, diese Syntax in R laufen zu lassen.
 - Wählen Sie das Arbeitsverzeichnis (working directory) geeignet.
 - Weisen Sie dem Objekt `matrikelnummer` Ihre eigene Matrikelnummer zu.
 - Ziehen Sie wie in der Vorlage eine Stichprobe vom Umfang 300 ohne Zurücklegen aus den Daten. Verwenden Sie dazu Ihre Matrikelnummer als Startwert.
 - Speichern Sie den resultierenden personalisierten Teil-Datensatz in den vier vorgegebenen Formaten. Verwenden Sie für die weiteren Aufgaben diesen Teildatensatz.
- Speichern Sie die Vorlage für die Abgabedatei „Abgabe_Einf_Software_Nachname_Vorname.docx“.
 - Ändern Sie den Dateinamen, indem Sie statt Nachname Ihren eigenen Nachnamen und statt Vorname Ihren eigenen Vornamen einsetzen, z.B. „Abgabe_Einf_Software_Oberhauser_Cornelia.docx“
 - Schreiben Sie in die ersten Zeilen des Dokumentes Ihren Nachnamen, Vornamen, Ihre Matrikelnummer und Ihre Email-Adresse. Geben Sie bitte außerdem an, ob Sie im Bachelor-Studiengang Statistik eingeschrieben sind und ob Sie eine ausführliche Rückmeldung zu Ihrer Abgabe wünschen.
 - Kopieren Sie unter Aufgabe 0) die R-Syntax, die Sie verwendet haben, um Ihren personalisierten Teil-Datensatz zu erzeugen.

Lösen Sie die folgenden Aufgaben in allen drei vorgestellten statistischen Softwarepaketen **R**, **SPSS** und **SAS**. Verwenden Sie für jedes Softwarepaket jeweils einen der gespeicherten, personalisierten Teil-Datensätze. (Verwenden Sie für jedes Softwarepaket eine Version des Datensatzes, die sich fehlerfrei einlesen lässt. Evtl. lassen sich alle Formate fehlerfrei in alle Softwarepakete einlesen, evtl. können bei einzelnen Formaten aber Probleme auftreten.)

Kopieren Sie für jedes Softwarepaket die Syntax und den Output (Tabellen, Graphiken) in Ihre Abgabedatei. Verwenden Sie dazu immer die Reihenfolge **R**, **SPSS** und **SAS**. Fangen Sie für jede Aufgabe eine neue Seite an. Falls irgendeine Form von Interpretation gefragt ist, schreiben Sie diese bitte ans Ende der jeweiligen Aufgabe.

Aufgabe 1:

Lesen Sie Ihren personalisierten Teil-Datensatz in das jeweilige Statistikprogramm ein. (Verwenden Sie dazu jeweils das Format des Teildatensatzes, welches Sie fehlerfrei in das jeweilige Programmpaket einlesen können.)

Aufgabe 2:

Erstellen Sie eine Häufigkeitstabelle von **school**. Berechnen Sie sowohl die absoluten als auch die relativen Häufigkeiten.

In welcher Schule wurden mehr Schüler befragt?

Aufgabe 3:

Erstellen Sie eine Häufigkeitstabelle von **reason**. Berechnen Sie sowohl die absoluten als auch die relativen Häufigkeiten.

Stellen Sie die absoluten Häufigkeiten von **reason** in einem Balkendiagramm dar.

Wie wichtig war der Ruf (reputation) der Schule bei der Schulwahl?

Aufgabe 4:

Erstellen Sie eine Kreuztabelle zwischen **school** und **reason**. Erstellen Sie wahlweise eine oder zwei Kreuztabellen mit den absoluten und den auf die Variable **school** bedingten relativen Häufigkeiten.

Stellen Sie die Variablen **school** und **reason** gemeinsam in einem Balkendiagramm dar. Verwenden Sie dazu ein auf 100% skaliertes Balkendiagramm der bedingten relativen Häufigkeiten.

In welcher Schule war ihr Ruf für die Wahl der besuchten Schule wichtiger?

Aufgabe 5:

Berechnen Sie für die drei Noten **G1**, **G2**, **G3** jeweils den Mittelwert, die Varianz, die Standardabweichung und die 5-Punkte-Zusammenfassung (d.h. Minimum, unteres Quartil, Median, oberes Quartil, Maximum).

In welchem Trimester waren die Noten am besten?

Aufgabe 6:

Erstellen Sie ein Histogramm der Schuljahres-Abschlussnote **G3**.

Wie lässt sich die Verteilung beschreiben?

Aufgabe 7:

Erstellen Sie einen Boxplot der Schuljahres-Abschlussnote **G3**.

In welchem Bereich liegen die Werte?

Aufgabe 8:

Erstellen Sie einen Boxplot der Schuljahres-Abschlussnote **G3** getrennt nach **reason**.

Unterscheiden sich die Noten in Abhängigkeit von dem Grund für die Wahl der Schule?

Aufgabe 9:

Zeichnen Sie jeweils ein Streudiagramm für den Zusammenhang zwischen **G1** und **G3** und für den Zusammenhang zwischen **G2** und **G3**.

Berechnen Sie die zugehörigen Korrelationen nach Spearman für diese beiden Zusammenhänge.

Hängt die Note am Ende des 1. Trimesters oder die Note am Ende des 2. Trimesters stärker mit der Schuljahres-Abschlussnote zusammen?

Geben Sie die von Ihnen umbenannte Abgabedatei „Abgabe_Einf_Software_Nachname_Vorname.docx“ bis spätestens **20. Januar 2017** über Moodle ab. (Wahlweise kann auch eine pdf-Datei abgegeben werden.)