**Grundlagen der Praktischen Informatik** **Sommersemester 2023**
Dr. Stefan Siemer Georg-August-Universität Göttingen
Timo Specht Institut für Informatik

# Übungsblatt 09

# Praktische Übung

**Abgabe der Prüfsumme bis Di., 18.06., 14 Uhr**
**Testat ab Di., 18.06., ab 18 Uhr**

**Hilfe** zum Bearbeiten der praktischen Übungen können Sie grundsätzlich jeden Tag in den Rechnerübungen bekommen. Die Testate finden ebenfalls in **Dreiergruppen** und Vierergruppen statt. Dabei sind die Gruppen identisch zu denen, die auch die theoretischen Aufgaben zusammen bearbeiten. In diesem Fall reserviert nur ein Gruppenmitglied einen Termin.

## Abgabe der Prüfsumme

- Für diese Übung müssen Sie **keine Prüfsumme** abgeben.
- Falls Sie eine Summe abgeben, übermitteln Sie die Prüfsumme mit dem Test *GdPI 09 - Testat*.

## Vorbereitung und Hinweise

- Dieses Übungsblatt enthält nur einen praktischen Teil, der auf das Python Modul Pandas fokussiert ist.
- Der Fokus dieses Zettels liegt auf dem Arbeiten mit Datensätzen, die Sie als `.csv` Dateien erhalten.
- Sie finden die Datensätze in StudIp unter *Grundlagen der Praktischen Informatik (Informatik II)* → *Dateien* → *Uebungen* → *uebung09-data*.
- Zum Bearbeiten des Übungszettels empfehlen wir Ihnen die Jupyter Cloud Umgebung der GWDG, da Sie dort direkt ein Python Notebook erstellen können und die nötigen Module bereits installiert sind. Des Weiteren wird eine übersichtliche Ausgabe geboten.
- Wichtig: Auf diesem Zettel können insgesamt 125 Punkte erzielt werden. Also 100 reguläre Punkte und 25 **extra** Punkte.

## Aufgabe 1 – Die Neuen (25 Punkte)

Sie sind zur Zeit angestellt an Hogwarts, der weltberühmten Schule für Zauberei und Hexerei. Ihre Aufgabe besteht darin, die Daten, die während eines Schuljahres anfallen, zu verwalten. Das heißt, Einpflegen der neuen Erstsemesterdaten, Auswerten der Daten zu den einzelnen Fächern und Berechnung der Punkteliste für den Hauspokal.

Ihre erste Aufgabe besteht darin den Datensatz der Erstsemester in das System einzupflegen. Folgende Aufgaben müssen dafür erledigt werden.

1. Laden Sie mit Hilfe des Moduls Pandas und der Methode `read_csv` den Datensatz `ErstSemester.csv` in einen Datenframe. (2 Punkte)

2. Lassen Sie die ersten 10 Zeilen anzeigen und machen Sie sich mit dem Datensatz vertraut. Nutzen Sie zum Ausgeben der Zeilen die Methode `head` von Pandas. (2 Punkte)

Bevor Sie nun die neuen Daten in die bereits vorhandenen Studierendendaten einfplegen, sollen folgende Aspekte von dem Datensatz untersucht werden.

1. Wie viele Erstsemester gibt es dieses Jahr und wie viel Prozent davon sind männlich bzw. weiblich? Benutzen Sie dafür **maximal drei Zeilen Code**. Geben Sie die Ergebnisse aus. (5 Punkte)

2. Wie viele Neuzugänge gibt es pro Haus und welches Haus hat die meisten Neuzugänge? Gehen Sie dabei wie folgt vor.

   a) Nutzen Sie die Methode `groupby`, um den Datenframe nach Häusern zu gruppieren. (2 Punkte)

   b) Iterieren Sie mittels einer `for` Schleife über die Gruppen und geben Sie jeweils den Namen der Gruppe und ihre zugehörige Größe aus, also wie viele Studierende in dieser Gruppe liegen. (2 Punkte)

   c) Geben Sie die Gruppe mit den meisten Studierenden aus. (2 Punkte)

Dieses Jahr will man zudem die Daten in ein neues Format bringen. Setzen Sie dafür folgende Anforderungen um.

1. Fassen Sie die Spalten, Vorname und Nachname zu einer neuen Spalte Name zusammen und löschen Sie anschließend diese beiden Spalten. Benutzen Sie dafür **nur eine Zeile Code**. (5 Punkte)
   Hinweis: Verwenden Sie die `assign` Methode in Kombination mit der `drop` Methode.

2. Extrahieren Sie aus den E-Mails die Benutzernamen und speichern Sie diese in einer neuen Spalte ab. Dabei ist der Benutzername alles bis zu @ in der E-Mail Zeichenkette. Benutzen Sie dafür **nur eine Zeile Code**. (5 Punkte)
   Hinweis: Verwenden Sie die `apply` Methode in Kombination mit `lambda` Ausdrücken.

## Aufgabe 2 – Jung trifft Alt (30 Punkte)

Im nächsten Schritt sollen nun die Erstsemesterdaten an die restlichen Studierendendaten angefügt werden.

1. Laden Sie mit Hilfe des Moduls Pandas und der Methode `read_csv` den Datensatz `StudierendenFDaten.csv` in einen Datenframe. (1 Punkt)

2. Geben Sie die letzten sieben Zeilen aus und machen Sie sich mit dem Datensatz vertraut. Nutzen Sie zum Ausgeben der Zeilen die Methode `tail` von Pandas. (1 Punkt)

Leider sind diese Daten auch noch im alten Format. Das heißt, Sie müssen erneut folgende Änderungen vornehmen.

1. Fassen Sie die Spalten, Vorname und Nachname zu einer neuen Spalte Name zusammen und löschen Sie anschließend diese beiden Spalten. Benutzen Sie dafür **nur eine Zeile Code**. (1 Punkt)
   Hinweis: Verwenden Sie die `assign` Methode in Kombination mit der `drop` Methode.

2. Extrahieren Sie aus den E-Mails die Benutzernamen und speichern Sie diese in einer neuen Spalte ab. Dabei ist der Benutzername alles bis zu @ in der E-Mail Zeichenkette. Gehen Sie wie folgt vorgehen.

   a) Schreiben Sie zunächst eine Methode `separate`. Diese bekommt eine Liste von Strings übergeben und einen Separator. Wird kein Separator übergeben, wird `","` als Separator verwendet. Die Methode soll jeden String in der Liste an dem Separator aufteilen und anschließend das Ergebnis zurückgeben. (5 Punkte)

   b) Verwenden Sie `separate`, um den Benutzernamen aus der E-Mail zu extrahieren. (5 Punkte)

3. Hängen Sie abschließend die Erstsemesterdaten an die Studierendendaten an. Nutzen Sie dafür die `concat` Methode von Pandas. Gehen Sie dabei wie folgt vor.

   a) Erklären Sie in eigenen Worten den Parameter `axis` der Methode `concat`. (2 Punkte)

   b) Für den Parameter `axis` gibt es zwei verschiedene Werte. Wenden Sie `concat` mit jedem dieser Werte an und speichern Sie das Ergebnis jeweils in einem separaten Datenframe. (2 Punkte)

   c) Erklären Sie die entstandenen Ergebnisse. Mit welchem sollten Sie weiterarbeiten? (2 Punkte)

   d) Ein weiterer Parameter von `concat` ist `ignore_index`. Dieser kann entweder auf `True` oder `False` gesetzt werden. Probieren Sie beides aus und erklären Sie den Unterschied. (2 Punkte)

Für die folgenden Aufgaben sollen Sie nun nur noch die kombinierten Daten verwenden. Zuerst sollen einige kleine Auswertungen erfolgen.

1. Welches Haus hat dieses Jahr prozentual gesehen den größten Zuwachs an Studierenden? (2 Punkte)

2. Ist der neue Jahrgang der größte Jahrgang? Wenn nicht, welcher Jahrgang hat mehr Studierende und wie viel Prozent sind dies? (2 Punkte)

3. Welches Haus hat den größten Männer- bzw. Frauenanteil ohne die Erstsemester und welches Haus hat den größten Männer- bzw. Frauenanteil mit den Erstsemestern? (5 Punkte)

## Aufgabe 3 – Der Hauspokal (45 Punkte)

Am Ende des Jahres erhalten Sie die Daten zu den einzelnen Fächern und wie viele Punkte die Studierenden jeweils in den Fächern erzielt haben. Auf Basis dieser Daten soll nun der Gewinner des Hauspokals ermittelt werden.

1. Laden Sie mit Hilfe des Moduls Pandas und der Methode `read_csv` die Datensätze `PflichtfächerPunkte.csv` und `WahlfächerPunkte.csv` jeweils in einen Datenframe. (1 Punkt)

2. Geben Sie jede zweite Zeile der Datensätze aus. Benutzen Sie dafür einmal die `loc` Methode und einmal die `iloc` Methode. Erklären Sie zudem den Unterschied zwischen `loc` und `iloc`. Wann sollte man `loc` und wann `iloc` verwenden? (3 Punkte)

3. Vereinigen Sie die beiden Datensätze mit den aktualisierten Daten der Studierenden. Nutzen Sie dafür die `merge` Methode von Pandas. Gehen Sie wie folgt vor.
   a) Erklären Sie in eigenen Worten den Parameter `how` der Methode `merge` (2 Punkte)
   b) Für den Parameter `how` gibt es fünf verschiedene Werte. Wenden Sie `merge` mit jedem dieser Werte an und speichern Sie das Ergebnis in einem separaten Datenframe. (5 Punkte)
   c) Erklären Sie die entstandenen Ergebnisse. Mit welchem sollten Sie weiterarbeiten? (5 Punkte)
   d) Erklären Sie den Unterschied zwischen `concat` und `merge`. (1 Punkt)

4. Welches Haus hat am meisten Punkte und bekommt somit den Hauspokal? (4 Punkte)

5. Berechnen Sie zudem pro Fach das Haus mit den meisten Punkten. (4 Punkte)

6. Ermitteln Sie pro Jahrgang den besten Studierenden. Also den Studierenden mit den meisten Punkten. (4 Punkte)

7. Berechnen Sie pro Fach den besten Jahrgang. Also den Jahrgang, der in diesem Fach die meisten Punkte erzielt hat. (4 Punkte)

Anschließend sollen Sie noch ein paar Daten zu den Punkteverteilungen in den Fächern sammeln.

1. Welches Fach hat die größte und welches Fach die geringste Standardabweichung? Geben Sie auch von jedem Fach die Mittelwerte aus. (4 Punkte)

2. Geben Sie die drei meistbesuchten Wahlfächer an. (4 Punkte)

3. Gibt es Studierende, die im dritten Schuljahr oder höher sind und kein Wahlfach belegt haben? (4 Punkte)

## Aufgabe 4 – Die beste Lehrperson (25 Punkte)

Abschließend erhalten Sie noch die Daten zu den Lehrveranstaltungsevaluationen.

1. Laden Sie mit Hilfe des Moduls Pandas und der Methode `read_csv` die Datensätze `VorlesungsBewertungen.csv` und `LehrpersonalDaten.csv` jeweils in einen Datenframe. (1 Punkt)

2. Leider sind die Bewertungen der Studierenden nicht anonymisiert. Das heißt, in einem ersten Schritt sollen die Daten anonymisiert werden, indem Sie die Namen durch die entsprechenden Matrikelnummern ersetzen. (5 Punkte)

3. Welches Fach wurde am besten bewertet und welche Lehrperson bekommt somit den Lehrpreis? (4 Punkte)

4. Welches Haus hat im Durchschnitt am besten und welches am schlechtesten bewertet? Gehen sie dabei wie folgt vor.

   a) Bilden Sie zunächst für jedes Haus pro Fach den Mittelwert. (2 Punkte)

   b) Nehmen Sie anschließend pro Haus den Mittelwert über die einzelnen Mittelwerte aus Aufgabenteil a). (2 Punkte)

   c) Welches Haus hat den maximalen und welches den minimalen Wert? (1 Punkt)

5. Welcher Jahrgang hat bei den Bewertungen jeweils die geringste Abweichung pro Fach?

   a) Berechnen Sie zuerst die Standardabweichung pro Fach für jedes Haus. (2 Punkte)

   b) Ermitteln Sie danach die Standardabweichung der Standardabweichungen pro Haus. (2 Punkte)

   c) Welches Haus hat den maximalen und welches den minimalen Wert? (1 Punkt)

6. Da das Schuljahr zu Ende ist, sollen Sie zudem alle Daten der Studierenden aus dem 7. Jahrgang entfernen. Das heißt, Sie sollen aus dem Datenframe von Aufgabe 2 alle Einträge zu den Studierenden aus dem 7. Jahrgang löschen. Dabei soll das Löschen *inplace* passieren. Das heißt, es soll kein neuer Datenframe erstellt werden nach dem Löschen. (3 Punkte)

7. Exportieren Sie den Datensatz aus Aufgabenteil 6) als `.csv`. (2 Punkte)