Dateien als Datenquellen

Handlungssituation

Eine der ersten Kunde der neu eingeführten Abteilung “Daten- und Prozessanalyse” der ChangeIT GmbH ist eine große berufsbildenden Schule.

Diese Schule möchte gerne die Leistungsdaten eines Jahrgangs ausgewertet haben. Diese Daten liegen sowohl als *csv*, *xml* und *json* vor.

Daten der Schülergruppe.

Die Daten der Schülergruppe liegen in drei unterschiedlichen Formaten vor.

CSV Darstellung

Eine Datei **moodle.csv** enthält die Daten in Form von CSV.

```
ID-Nummer;Abteilung;Klassenarbeit;K1;K2;K3;K4;K5;K6;K7
1;J;100.00 %;90.00 %;100.00 %;92.86 %;100.00 %;100.00 %;100.00 %;100.00 %
2;J;80.62 %;70.00 %;100.00 %;64.29 %;72.22 %;77.27 %;83.33 %;93.58 %
3;J;95.07 %;95.00 %;100.00 %;100.00 %;100.00 %;100.00 %;100.00 %;100.00 %
```

JSON Darstellung

Eine Weitere Datei **Moodle.json** enthält die Daten in JSON Form.

```
[
 {
   "ID-Nummer": 1,
   "Abteilung": "J",
   "Klassenarbeit": "100.00 %",
   "K1": "90.00 %",
   "K2": "100.00 %",
   "K3": "92.86 %",
   "K4": "100.00 %",
   "K5": "100.00 %",
   "K6": "100.00 %",
   "K7": "100.00 %"
 }
]
```

XML Darstellung

Eine weitere Datei **Moodle.xml** erhält die Daten in Form einer XML Datei.

```
<root>
  <row>
    <ID-Nummer>1</ID-Nummer>
    <Abteilung>J</Abteilung>
    <Klassenarbeit>100.00 %</Klassenarbeit>
    <K1>90.00 %</K1>
    <K2>100.00 %</K2>
    <K3>92.86 %</K3>
    <K4>100.00 %</K4>
    <K5>100.00 %</K5>
    <K6>100.00 %</K6>
```

```
    <K7>100.00 %</K7>
  </row>
</root>
```

Attribute des Datensatzes

Die einzelnen Attribute des Datensatzes haben dabei folgende Bedeutung.

- ID-Nummer: Eindeutige ID eines Schülers
- Abteilung: Jeweilige Klasse des Schülers (J-N)
- Klassenarbeit: Ergebnis der Klassenarbeit
- K1 bis K7: Ergebnisse in einzelnen Kapitel-Tests.

Arbeitsprozess

Im weiteren Verlauf der Kurses werden wir zunächst folgenden Arbeitsprozess anwenden.

Figure 1: Arbeitsprouess

- Wir werden zunächst die Rohdaten aus einer Datenquelle einlesen. In unserem Fall ist das aktuell eine Datei.
- Anschließend werden wir die Daten anpassen, bereinigen oder vorauswerten.
- Abschließend werden wir das Ergebnis ansprechend für den Kunden visualisieren.

Einlesen der Dateien

Hinweise

> Für die Analyse und Visualisierung der Daten verwenden wir das Python Paket **pandas** verwenden. Wenn Sie diese Pakete noch nicht installiert haben, so holen Sie dieses bitte nach.

Ein Paket in Python installiert man über den Python Paket Manger PIP wie folgt:

```
pip install {Name des Paketes}
```

Aufgabe: Daten einlesen

Entscheiden Sie sich für ein Dateiformat welches Sie bearbeiten wollen? Bilden Sie dazu Gruppen für die drei genanten Dateitypen und schreiben Sie ein Python Programm welches die jeweilige Datei einließt. Dabei sollten die Daten gleich in einem DataFrame (Datenformat für die Auswertung mit **pandas**) eingelesen werden. Geben Sie nach dem Einlesen des Datensatzes das Dataframe im Jupyter Notebook aus. Die Ausgabe sollte der folgenden Abbildung entsprechen.

> Diskutieren Sie in ihrer Gruppe, ob die Daten bereits so wie Sie vorliegen verwendet werden können, oder ob noch Bereinigungen der Daten notwendig sind?

Bereinigen der Dateien

Leider sind die Daten für die weitere Bearbeitung nicht geeignet und müssen entsprechend angepasst werden. Folgende Probleme wurden in den Daten festgestellt.

	ID-Nummer	Abteilung	Klassenarbeit	K1	K2	K3	K4	K5	K6	K7
0	1	J	100.00 %	90.00 %	100.00 %	92.86 %	100.00 %	100.00 %	100.00 %	100.00 %
1	2	J	80.62 %	70.00 %	100.00 %	64.29 %	72.22 %	77.27 %	83.33 %	93.58 %
2	3	J	95.07 %	95.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
3	4	J	85.80 %	50.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
4	5	J	94.93 %	90.00 %	100.00 %	100.00 %	88.89 %	95.45 %	100.00 %	91.30 %
...	...	...	...	...	...	...	...	...	...	...
119	120	N	89.31 %	75.00 %	100.00 %	100.00 %	94.44 %	100.00 %	100.00 %	100.00 %
120	121	N	88.62 %	90.00 %	100.00 %	71.43 %	100.00 %	95.45 %	100.00 %	52.38 %
121	122	N	60.33 %	100.00 %	92.86 %	92.86 %	100.00 %	95.45 %	100.00 %	-
122	123	N	79.46 %	90.00 %	85.71 %	57.14 %	69.44 %	95.45 %	86.11 %	48.03 %
123	124	N	68.26 %	45.00 %	89.29 %	67.86 %	80.56 %	63.64 %	-	37.27 %

124 rows × 10 columns

Figure 2: Dataframe Rohdaten

- Die Notenwerte liegen in den Daten als Zeichenketten vor, sinnvoller wäre es hier, wenn die Daten als *float* vorliegen würden.
- Wenn ein Schüler einen Kapitel-Test nicht mitgeschrieben hat, so ist der Datensatz mit '-' gekennzeichnet. Sinnvoller wäre es, wenn hier der Wert 0.0 eingetragen wäre.

Aufgabe: Daten umwandeln

Wandeln Sie die Werte des DataFrames entsprechend der oben durchgeführten Überlegungen um und geben Sie anschließend das DataFrame Objekt im Jupyter Notebook um und kontrollieren Sie die Ausgabe.

Erste Statistische Grunddaten ermitteln

Wir wollen nun die Daten einer ersten statistischen Untersuchung unterziehen.

Das arithmetische Mittel eines Datenmenge berechnet sich wie folgt:

Die Standardabweichung einer Datenmenge X berechnet sich wie folgt:

TODO hier ergänzen

$\sigma = \sum_{i=0}^{i=N}(Xi)$

statistische Grunddaten ermitteln

Ermitteln Sie mit Hilfe von **pandas** folgende Daten.

- Durchschnittswert (in %) der Klassenarbeit
- Standradabweichung der Klassenarbeit
- Durchschnittswert (in %) der Klassenarbeit der Gruppe "J"
- Notenspiegel der Klassenarbeit (also wie viele Schüler haben ein "1", wie viele Schüler eine "2" usw.)

Hinweis: Nutzen Sie zum ermitteln der Daten die Dokumentation zum pandas Dataframe.

Daten visualisieren

Der Auftraggeber (die Schule) möchte die Daten zur weiteren Verarbeitung in Dokumenten visualisiert haben. Zur Visualisierung von Daten nutzen wir das python Paket **plot** aus dem **pandas** Paket.

Für die folgenden Aufgaben stellt Ihnen der Leiter der Abteilung Datenanalyse folgenden Beispielcode zur Verfügung.

Aufgabe 1 Daten visualisieren

Der Auftraggeber würde gerne das Ergebnis der Klassenarbeit in Form eines Tortendiagramms dargestellt haben. Nutzen Sie die eingelesenen Daten um dieses Tortendiagramm darzustellen.

Hinweis: In dem unten dargestellten Beispiel werden Werte in Form eines Tortendiagramms dargestellt.

```
import pandas as pd

data = pd.DataFrame([1,2,3,4,5,6])
pieimage = data.plot.pie(subplots=True)
pieimage
```

Aufgabe 2 Daten visualisieren

Es stellt dich die Frage, ob Schüler die im arithmetischen Mittel in den Kapitel-Tests eine gute Note schreiben auch in der Klassenarbeit eine gute Note schreiben. Versuchen Sie auf diese Fragestellung eine Antwort zu finden und visualisieren Sie das Ergebnis möglichst sinnvoll für den Auftragsgeber.

Hinweis: In dem unten dargestellten Beispiel werden mit Hilfe von **pandas** Punkte mit (x/y) Werten in einem Koordinatensystem dargestellt.

```
import pandas as pd

data = pd.DataFrame({"x":[1,2,3,4,5,6],"y":[6,5,4,3,2,1]})

pltimage = data.plot.scatter(x="x",y="y")
pltimage
```