

```

1  ### md
2  # Statistik-Labor Testat Aufgabe 1
3  ***Hinweise:***
4  - Bitte überprüfen Sie Ihre Resultate vor der Abgabe
  Ihre Ergebnisse mit den Teilergebnissen aus der Datei
    ***Teilergebnisse_xy.txt***. In der Datei bekommen
  Sie jeweils die ersten beiden Nachkommastellen des
  Durchschnittsalters.
5  - Die Unterlagen sind im **pdf-Format** in Moodle
  hochgeladen abzugeben. Richtige Lösungen werden nicht
  mehr an Sie zurückgegeben. Eine Abgabe per Email ist
  nicht möglich.
6  - Bitte füllen Sie das jeweilige **Deckblatt** mit
  aus und geben es mit Ihrer Lösung zusammen ab.
7
8  ***Aufgabenstellung:***<br>
9  Sie sollen für die Personalabteilung der NANUAG die
  Altersverteilung der Beschäftigten analysieren. <br>
10 In der Datei **sr_aufg_1_xy.txt** im Unterverzeichnis
    **Endziffer_xy** (*„xy“ steht für die beiden
  Endziffern Ihrer Matrikelnummer*) finden Sie die
  dafür erforderlichen Daten. <br>
11 Diese Datei enthält folgende fünf Felder, die durch
  Semikolons getrennt sind: Personalnummer, Anrede („
  Herr“ oder „Frau“), Vorname, Nachname, Geburtsdatum.
12
13 ## 1. Daten einlesen
14 Lesen Sie Daten in **Python** bzw. ein **Jupyter
  Notebook** ein. Speichern Sie sie in einem **Pandas
  DataFrame** ab.
15
16 ***Tipp:*** Bitte beachten Sie bei der Abgabe, dass
  Sie auf keinen Fall die Liste mit den 100 Datensätzen
  ausgeben! Verwenden Sie dafür zum Beispiel die `
  dataframe.head(x)`-Funktion.
17 ###
18 import pandas as pd
19
20 #ToDo: 1. Ersetzen Sie im Dateinamen xy durch
  entsprechende Matrikelnummer
21 #      2. Lassen Sie sich alle Daten anzeigen, um "

```

```

21 fehlerhafte" Daten zu erkennen
22
23 data = pd.read_csv('datasets/sr_aufg_1_13.txt', sep=
    ';',
24                     encoding='unicode_escape', na_values='. '
    , header = None,
25                     names=['Pers.Nr.', 'Anrede', 'Vorname', '
    Nachname', 'Geburtsdatum'])
26 data
27 ### md
28 ## 2. Daten bereinigen
29 Führen Sie eine Datenbereinigung durch. <br>
30 Listen Sie in einer Tabelle auf, welche (Daten-)Sätze
    Sie gelöscht und welche Sätze Sie wie korrigiert
    haben.
31
32 ***Tipp:*** Sie können hierzu die Python Funktionen `
    dataframe.drop()` und `dataframe.at()` verwenden. <br>
    >
33 ***Beispiellösung:***
34 ![Beispiellösung](https://drive.google.com/uc?id=
    1z5Gydg1HzoxZLo428862fsgsKyaHKKm9)
35 ###
36 #Filter invalid entries in 'Geburstdatum'
37 data_korrektur = pd.DataFrame(columns=['Pers.Nr.', '
    Anrede', 'Vorname', 'Nachname', 'Geburtsdatum', '
    Datensatz?', 'Typ des Fehlers'])
38
39 #Korrekturtabelle füllen
40 data_korrektur.loc[0] = data.loc[8]
41 data_korrektur.at[0, 'Datensatz?'] = 'Löschen'
42 data_korrektur.at[0, 'Typ des Fehlers'] = 'Jahr 25408
    ?'
43
44 data_korrektur.loc[1] = data.loc[25]
45
46 data_korrektur
47
48 ### md
49 ## 3. Alter bestimmen
50 Berechnen Sie das Alter der Beschäftigten zum

```

```

50 Stichtag 31.12.2005. Mit „Alter“ ist hier die Anzahl
    der vollendeten Lebensjahre gemeint. <br>
51 *Beispiel:* Wer am 30.06.1975 geborgen ist, ist zum
    Stichtag 30 Jahre alt. <br>
52 (Niemand würde von sich sagen, er wäre „30,5 Jahre
    alt“.) Folglich wird diese Person in der Alterklasse
    21 - 30 Jahre mitgezählt.
53
54 ***Tipp:*** In Python ist hierzu die Pandas-
    Bibliothek mit ihre Funktion `pandas.to_datetime()`,
    sowie `pandas.year` nützlich.
55 #%%
56
57 #%% md
58 ## 4. Säulendiagramm mit Altersverteilung
59 Stellen Sie die Altersverteilung der Beschäftigten in
    einem gruppierten Säulendiagramm dar, in dem Männer
    und Frauen getrennt ausgewiesen werden.<br>
60 Es sollen die Altersklassen bis (einschließlich) 20
    Jahre, 21 - 30 Jahre, 31 - 40 Jahre, 41 - 50 Jahre
    und über 50 Jahre gewählt werden.
61
62 ***Tipp:*** Sie benötigen hierfür die Python
    Bibliothek Matplotlib und NumPy sowie aus 
    Pandas die Funktion `value_counts()`. <br>
63 ***Beispiellösung:***
64 <div>
65 
66 </div>
67 #%%
68
69 #%% md
70 ## 5. Kreis-/Tortendiagramm der Altersverteilung
71 Stellen Sie die Altersverteilung der männlichen
    Beschäftigten und der weiblichen Beschäftigten in je
    einem Kreisdiagramm („Tortendiagramm“) dar. <br>
72 Die Kreissektoren sind mit den zugehörigen
    Prozentzahlen zu beschriften.
73
74 ***Beispiellösung:***

```

```
75 <div>
76 
77 </div>
78 #%%
79
80 #%% md
81 ## 6. Kennzahlen
82 Berechnen Sie Mittelwert, empirische Varianz
    empirische Standardabweichung, Median
    und Spannweite der Alterszahlen:
83 - für alle Beschäftigten und
84 - getrennt für Männer und Frauen.
85
86 ***Beispiellösung:***
87 <div>
88 
89 </div>
90 #%%
91
92 #%%
93
```