






Der DataScience – Workflow in Python

	1.)	2.)	3.)
Arbeitsschritte	<ul style="list-style-type: none"> Daten aus einer csv-Datei oder Excel-Datei einlesen Daten sortieren und filtern 	<ul style="list-style-type: none"> Mathematische Berechnungen mit Daten durchführen 	<ul style="list-style-type: none"> Daten visuell darstellen
Tool (Modul)	  	 NumPy	
Wie Modul einbinden?	<code>import pandas as pd</code>	<code>import numpy as np</code>	<code>%matplotlib inline</code> <code>import matplotlib.pyplot as plt</code>
Relevante Datenstruktur	DataFrame	Array	

1.1.1 csv

Mit dem csv - Modul lassen sich CSV-Daten komfortabel laden (CSV = comma separated values)

Modul einbinden

```
In [2]: import csv
```

Anwendung

```
In [3]: with open("./resources/datei.csv", encoding="utf-8") as file:
        csv_file = csv.reader(file, delimiter=",")
        for line in csv_file:
            print(line)
```

Anstatt:

```
# with open("datei.csv") as file:
#     for line in file:
#         data = line.strip().split(",")
#         print(data)
```

```
['Name', 'Telefonnummer', 'Land']
['Müller', '+49123456789', 'Deutschland']
['Mustermann', '+3612345678', 'Ungarn']
```

Weitere Infos: <https://docs.python.org/3/library/csv.html>

1.1.2 pandas

Essentielles Modul zur Datenanalyse mit Python, auch wegen der DataFrame - Struktur.

Modul einbinden

```
In [4]: import pandas as pd # Umbenennung ist Konvention
```

Anwendung

```
In [5]: # CSV-Datei als DataFrame einlesen
df = pd.read_csv("../data/astronauts.csv", delimiter=",")
df[["Name", "Year", "Gender"]].head()
```

```
Out[5]:
```

	Name	Year	Gender
0	Joseph M. Acaba	2004.0	Male
1	Loren W. Acton	NaN	Male
2	James C. Adamson	1984.0	Male
3	Thomas D. Akers	1987.0	Male
4	Buzz Aldrin	1963.0	Male

```
In [6]: # DataFrame nach Frauen filtern, die vor 2000 auf Mission waren
```

```
df2 = df[df["Year"] < 2000]
df3 = df2[df2["Gender"] == "Female"]
df3[["Name", "Year", "Gender"]].head()
```

```
Out[6]:
```

	Name	Year	Gender
19	Ellen S. Baker	1984.0	Female
50	Yvonne D. Cagle	1996.0	Female
52	Tracy E. Caldwell (Dyson)	1998.0	Female
67	Kalpana Chawla	1995.0	Female
70	Laurel B. Clark	1996.0	Female

Weitere Infos: <https://pandas.pydata.org/pandas-docs/stable/tutorials.html>

1.1.3 NumPy

NumPy vereinfacht wissenschaftliches Rechnen, vor allem durch die Array - Datenstruktur.

Modul einbinden

```
In [7]: import numpy as np # Umbenennung ist Konvention
```

Anwendung

```
In [8]: x = np.arange(10) * 3  
        y = np.zeros(10) + 4
```

```
        z = x + y  
        z = z.reshape(5,2)
```

```
        print(z)  
        print(type(z))
```

```
[[ 4.  7.]  
 [10. 13.]  
 [16. 19.]  
 [22. 25.]  
 [28. 31.]]  
<class 'numpy.ndarray'>
```

Weitere Infos: <https://docs.scipy.org/doc/numpy-1.13.0/user/index.html>

1.1.4 matplotlib

Ermöglicht das Visualisieren von Daten.

Modul einbinden

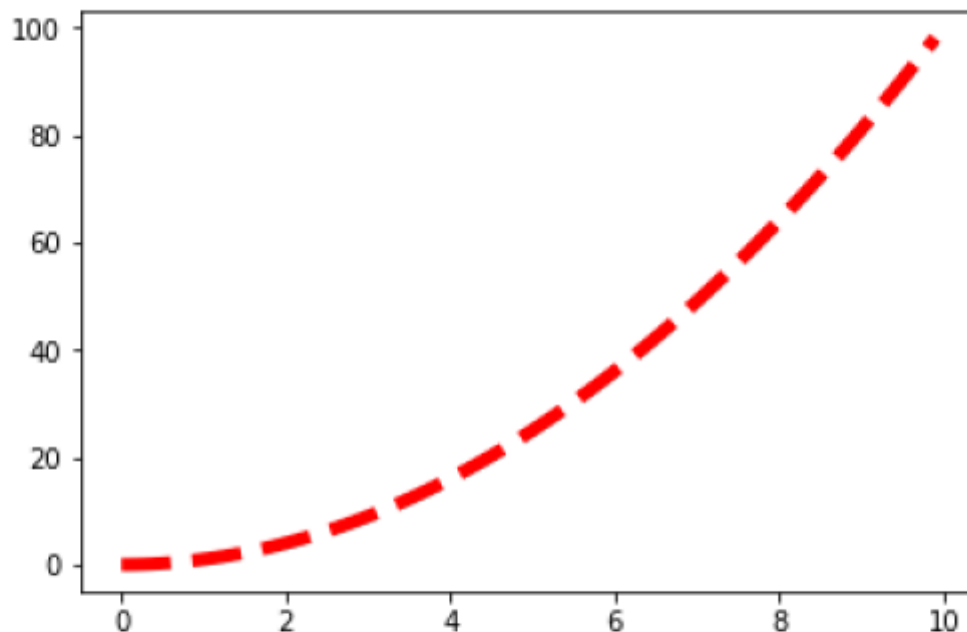
```
In [9]: # damit die Grafiken innerhalb des Notebooks angezeigt werden
        %matplotlib inline

import matplotlib.pyplot as plt # Umbenennung ist Konvention
```

Typische Anwendung

```
In [10]: xs = [x / 10 for x in range(0, 100)]
        ys = [x * x for x in xs]

        # Wir plotten einen Graphen durch die gegebenen Punkte
        plt.plot(xs, ys, color="r", linewidth=5, linestyle="dashed")
        plt.show()
```



Weitere Infos: <https://matplotlib.org/tutorials/index.html>
