Einstieg in Pandas

große Datensätze analysieren mit Pandas

Was ist Pandas?

- Bibliothek auf Basis von `NumPy`
- Stellt Datenstrukturen und Tools zur Analyse großer Datenmenge zur Verfügung
- Vorteile ggü. Arbeit mit Excel:
  - deutlich performanter
  - Open-Source, große Community

Datenstrukturen in Pandas

- Pandas arbeitet mit 2 Datenstrukturen: `Series` und `DataFrames`
  - `Series`: 1-dimensional mit Label (vgl. Excel-Spalte/Zeile)
  - `DataFrames`: 2-dimensional mit N Spalten (vgl. Excel-Tabelle)

Daten in Pandas importieren

- Pandas bietet Funktionen um die meisten Datenformate zu importieren
  - Heute: .csv-Daten `read_csv(pfad-zur-datei)`

In [1]:
```
import pandas
csv_daten = pandas.read_csv('../datasets/kundendaten_mock.csv')
csv_daten
```

Out[1]:

	Unnamed: 0	id	Vorname	Nachname	
0	0	1	Owen	Lambol	olambol0@oaic.gov.au
1	1	2	Cullin	Goodwell	cgoodwell1@shinystat.com
2	2	3	Weber	Majury	wmajury2@techcrunch.com
3	3	4	Felipe	Prendergrast	fprendergrast3@hostgator.com
4	4	5	Helyn	Frounks	hfrounks4@123-reg.co.uk
5	5	6	Ode	Lambal	olambal5@unblog.fr
6	6	7	Estell	Grandison	egrandison6@sciencedaily.com
7	7	8	Jodee	Sulman	jsulman7@oracle.com
8	8	9	Graig	McPhater	gmcphater8@ucoz.ru
9	9	10	Tammy	Faraday	tfaraday9@1und1.de
10	10	11	Lynnette	Ohanessian	lohanessiana@skyrock.com
11	11	12	Thorvald	Cobbing	tcobbingb@angelfire.com
12	12	13	Etienne	Pasquale	epasqualec@rambler.ru
13	13	14	Marji	Camerana	mcameranad@bloglovin.com
14	14	15	Dorisa	Knyvett	dknyvette@4shared.com

	Unnamed: 0	id	Vorname	Nachname	
15	15	16	Wait	Guess	wguessf@domainmarket.com
16	16	17	Lynne	Keedwell	lkeedwellg@biblegateway.com
17	17	18	Ermin	Kennally	ekennallyh@blogspot.com
18	18	19	Borg	Feedome	bfeedomei@arstechnica.com
19	19	20	Bogey	Penburton	bpenburtonj@home.pl
20	20	21	Eduard	Truse	etrusek@com.com
21	21	22	Kasper	Verchambre	kverchambrel@scientificamerica
22	22	23	Jeramey	Kegan	jkeganm@tripod.com
23	23	24	Terrijo	MacGilrewy	tmacgilrewyn@yelp.com
24	24	25	Karoly	Spelsbury	kspelsburyo@flavors.me
25	25	26	Elia	Coulman	ecoulmanp@reuters.com
26	26	27	Madelin	Frankcomb	mfrankcombq@cargocollective.c
27	27	28	Ivor	McDavitt	imcdavittr@harvard.edu
28	28	29	Celine	Colpus	ccolpuss@gnu.org
29	29	30	Roderich	Ellgood	rellgoodt@telegraph.co.uk
...	...	...	...	...	...
70	70	71	Dario	Massard	dmassard1y@gnu.org
71	71	72	Alicia	Huckin	ahuckin1z@jalbum.net

	Unnamed: 0	id	Vorname	Nachname	
72	72	73	Bea	Parfrey	bparfrey20@wordpress.org
73	73	74	Rudiger	Bircher	rbircher21@hibu.com
74	74	75	Caye	Gillyett	cgillyett22@pen.io
75	75	76	Abbot	Flecknell	aflecknell23@ca.gov
76	76	77	Lalo	Riccardini	lriccardini24@utexas.edu
77	77	78	Nathan	Carson	ncarson25@vk.com
78	78	79	Stacee	Mullenger	smullenger26@facebook.com
79	79	80	Pearline	Deverose	pdeverose27@spotify.com
80	80	81	Rudie	Froud	rfroud28@barnesandnoble.com
81	81	82	Haleigh	Huyge	hhuyge29@123-reg.co.uk
82	82	83	Iolanthe	Labone	ilabone2a@netlog.com
83	83	84	Robena	Thomazet	rthomazet2b@opera.com
84	84	85	Derby	Hingeley	dhingeley2c@alexa.com
85	85	86	Mireille	Scothorn	mscothorn2d@toplist.cz
86	86	87	Cherin	Bariball	cbariball2e@webnode.com
87	87	88	Lazar	Ealles	lealles2f@about.com
88	88	89	Sidonia	Gilgryst	sgilgryst2g@clickbank.net
89	89	90	Abby	Pietruszewicz	apietruszewicz2h@over-blog.com

Einen Überblick verschaffen

- `df.head(n)` zeigt die ersten n Zeilen eines `DataFrames` an , analog `df.tail(n)`
- `df.sample(n)` Stichprobe der Größe n entnehmen
- `df.sort_values(by=)` sortiert Tabelle nach gegebener Spalte
- `df.describe()` zeigt verschiedene Statistiken (Median, usw.) zu den Daten an

In [2]:
```
#csv_daten.head(10)
#csv_daten.tail(10)
#csv_daten.sort_values(by='Alter')
pandas.set_option('display.float_format', lambda x: '%.2f' % x)

#csv_daten.describe()
csv_daten.sample(n=20)
```

Out[2]:

	Unnamed: 0	id	Vorname	Nachname	Email	Ges
87	87	88	Lazar	Ealles	lealles2f@about.com	M
76	76	77	Lalo	Riccardini	lriccardini24@utexas.edu	M
99	99	100	Blondell	Meins	bmeins2r@gravatar.com	F
12	12	13	Etienne	Pasquale	epasqualec@rambler.ru	M
17	17	18	Ermin	Kennally	ekennallyh@blogspot.com	M
86	86	87	Cherin	Bariball	cbariball2e@webnode.com	F
37	37	38	Arliene	Poupard	apoupard11@digg.com	F
34	34	35	Siana	Salazar	ssalazary@nbcnews.com	F
27	27	28	Ivor	McDavitt	imcdavittr@harvard.edu	M
82	82	83	Iolanthe	Labone	ilabone2a@netlog.com	F
83	83	84	Robena	Thomazet	rthomazet2b@opera.com	F
77	77	78	Nathan	Carson	ncarson25@vk.com	M
46	46	47	Edsel	Maryan	emaryan1a@rediff.com	M

Daten selektieren

- nach Zeile `df.loc[1]` bzw. Abschnitt `df.loc[:10]`
- nach Spalte `df['Alter']` bzw. `df['Alter'][10:50]`
- nach Eigenschaften `df[df['Geschlecht'] ==M]` oder `df.query()`

In [3]:
```
csv_daten
#csv_daten.loc[10]
#csv_daten.loc[:10]
#csv_daten['Alter'][10:50]
#csv_daten['Email'].sample(10)
#csv_daten[csv_daten['Geschlecht'] == 'M']
csv_daten.query("Alter > 20")
```

Out[3]:

	Unnamed: 0	id	Vorname	Nachname	
0	0	1	Owen	Lambol	olambol0@oaic.gov.au
1	1	2	Cullin	Goodwell	cgoodwell1@shinystat.com
3	3	4	Felipe	Prendergrast	fprendergrast3@hostgator.com
4	4	5	Helyn	Frounks	hfrounks4@123-reg.co.uk
5	5	6	Ode	Lambal	olambal5@unblog.fr
6	6	7	Estell	Grandison	egrandison6@sciencedaily.com
7	7	8	Jodee	Sulman	jsulman7@oracle.com
8	8	9	Graig	McPhater	gmcphater8@ucoz.ru
9	9	10	Tammy	Faraday	tfaraday9@1und1.de
11	11	12	Thorvald	Cobbing	tcobbingb@angelfire.com
12	12	13	Etienne	Pasquale	epasqualec@rambler.ru
13	13	14	Marji	Camerana	mcameranad@bloglovin.com
14	14	15	Dorisa	Knyvett	dknyvette@4shared.com

	Unnamed: 0	id	Vorname	Nachname	
15	15	16	Wait	Guess	wguessf@domainmarket.com
16	16	17	Lynne	Keedwell	lkeedwellg@biblegateway.com
17	17	18	Ermin	Kennally	ekennallyh@blogspot.com
18	18	19	Borg	Feedome	bfeedomei@arstechnica.com
19	19	20	Bogey	Penburton	bpenburtonj@home.pl
20	20	21	Eduard	Truse	etrusek@com.com
21	21	22	Kasper	Verchambre	kverchambrel@scientificamerica
22	22	23	Jeramey	Kegan	jkeganm@tripod.com
23	23	24	Terrijo	MacGilrewy	tmacgilrewyn@yelp.com
24	24	25	Karoly	Spelsbury	kspelsburyo@flavors.me
25	25	26	Elia	Coulman	ecoulmanp@reuters.com
26	26	27	Madelin	Frankcomb	mfrankcombq@cargocollective.c
27	27	28	Ivor	McDavitt	imcdavittr@harvard.edu
28	28	29	Celine	Colpus	ccolpuss@gnu.org
29	29	30	Roderich	Ellgood	rellgoodt@telegraph.co.uk
30	30	31	Court	Iggulden	cigguldenu@msu.edu
31	31	32	Tadeo	Scowcraft	tscowcraftv@virginia.edu
...	...	...	...	...	...

	Unnamed: 0	id	Vorname	Nachname	
68	68	69	Tomkin	Samter	tsamter1w@miibeian.gov.cn
69	69	70	Lizbeth	Vannoort	lvannoort1x@flickr.com
70	70	71	Dario	Massard	dmassard1y@gnu.org
71	71	72	Alicia	Huckin	ahuckin1z@jalbum.net
72	72	73	Bea	Parfrey	bparfrey20@wordpress.org
73	73	74	Rudiger	Bircher	rbircher21@hibu.com
74	74	75	Caye	Gillyett	cgillyett22@pen.io
75	75	76	Abbot	Flecknell	aflecknell23@ca.gov
76	76	77	Lalo	Riccardini	lriccardini24@utexas.edu
77	77	78	Nathan	Carson	ncarson25@vk.com
78	78	79	Stacee	Mullenger	smullenger26@facebook.com
79	79	80	Pearline	Deverose	pdeverose27@spotify.com
81	81	82	Haleigh	Huyge	hhuyge29@123-reg.co.uk
83	83	84	Robena	Thomazet	rthomazet2b@opera.com
84	84	85	Derby	Hingeley	dhingeley2c@alexa.com
85	85	86	Mireille	Scothorn	mscothorn2d@toplist.cz
86	86	87	Cherin	Bariball	cbariball2e@webnode.com
87	87	88	Lazar	Ealles	lealles2f@about.com

Abschlussaufgabe: Explorativ Datensätze analysieren

- Als Abschlussaufgabe für den heutigen Tag haben wir 2 Datensätze vorbereitet
  - "Wine review"-Datensatz
  - Kickstarter-Datensatz
- Entnommen von der Plattform Kaggle (http://www.kaggle.com)
  - https://www.kaggle.com/kemical/kickstarter-projects (https://www.kaggle.com/kemical/kickstarter-projects)
  - https://www.kaggle.com/zynicide/wine-reviews (https://www.kaggle.com/zynicide/wine-reviews)

Kickstarter-Datensatz

- Crowdfunding-Plattform Kickstarter (http://kickstarter.com)
- Metadaten wie:
  - Projektname, Projektkategorie
  - Erfragte Summe /Erhaltene Summe

In [7]:
```
import pandas
kickstarter = pandas.read_csv(r'../datasets/kickstarter-projects/ks-projects-201801.csv'
)

kickstarter.head()
```

Out[7]:

	ID	name	category	main_category	currency	deadline
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16

Wine-Review Datensatz

- Datensatz mit Wein-Bewertungen von der Plattform WineEnthusiast (http://www.winemag.com/?s=&drink_type=wine)
- Metadaten wie:
  - Beschreibung des Sommeliers
  - Herkunftsland/-region
  - Winzer
  - Bewertung auf einer Punkte-Skala

In [8]:
```
wine_reviews = pandas.read_csv(r"..\datasets\wine-reviews\winemag-data_first150k.csv")

wine_reviews.head()
```

Out[8]:

	Unnamed: 0	country	description	designation	points	price	province
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.00	California
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.00	Northern Spain
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.00	California

Aufgabe:

1. Importieren sie die csv-Daten in ihr Jupyter-Notebook
2. Verschaffen sie sich einen Überblick über den Datensatz (`head()`, `describe()`, `sample(n)`)
3. Überlegen sie sich Hypothesen , die sie untersuchen möchten
4. Untersuchen sie die Hypothesen (+ evtl. Visualisierung)