

Tag 8

Inhaltsverzeichnis

- Bio- und konventionellen BE, JU und NE Landwirtschaftsbetriebe vergleichen
- Umgang mit fehlenden Daten
- Umgang mit Ausreißern
- Daten mit dem AutorBuch Beispiel "joinen"
- Übungen
- BYOQ

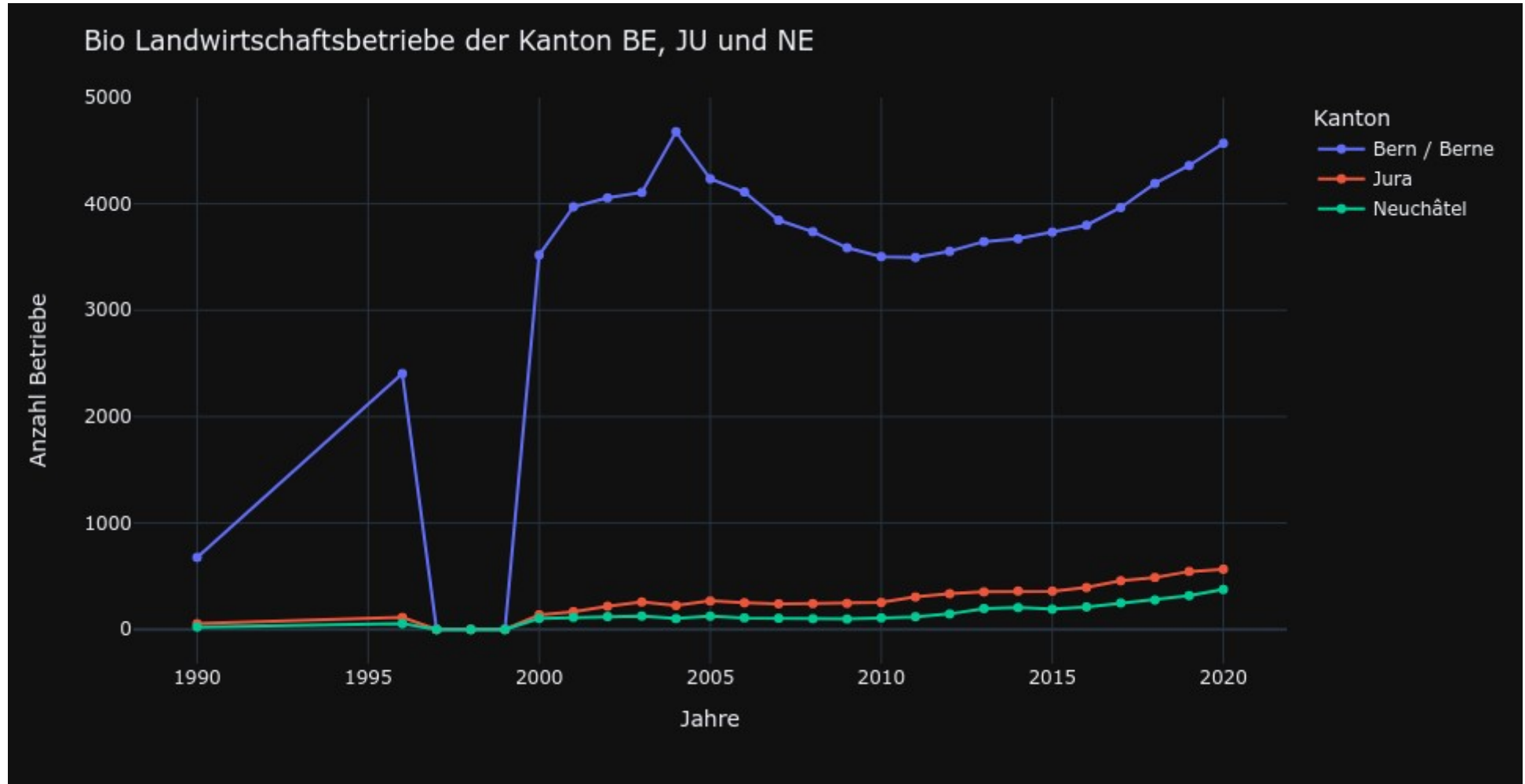
Vergleich der Kantone BE, JU und NE Datenbereinigung

Siehe Notebook

`Notebook_ITP_DM0D8_Bio-LWBetriebe-BEJUNE-V?.ipynb`

Vergleich der Kantone BE, JU und NE

Grafische Darstellung



Umgang mit fehlenden Daten

Unterschiedliche Lösungen

Variable	Kanton	Betriebssystem	1990	2000	2010	2020
Beschäftigte Total	Zürich	Konventionelle Betriebe	18832	13832	10907	9227
Beschäftigte Total	Bern / Berne	Konventionelle Betriebe	52640	38082		27658
Beschäftigte Total	Luzern	Konventionelle Betriebe	19821	15921	13638	12004

```
[2]: missingdata_df = pd.read_csv('../Data/missingdata.csv',
    sep=';', header=1, usecols=['Kanton', '1990', '2000', '2010', '2020'], encoding = 'ISO-8859-1')
print('Shape: ' + str(missingdata_df.shape))
missingdata_df.head()
```

Shape: (3, 5)

```
[2]:
```

	Kanton	1990	2000	2010	2020
0	Zürich	18832	13832	10907.0	9227
1	Bern / Berne	52640	38082	NaN	27658
2	Luzern	19821	15921	13638.0	12004

NaN == Not a Number (float)

- Problem: Was machen wir, wenn Daten fehlen?
- Pandas bietet unterschiedliche Lösungen
 - 1) Sie zuerst mit `isnull()` entdecken
 - 2) Datensatz mit `dropna()` löschen (Zeile oder Kolonne)
 - 3) NaN-Wert mit einem bestimmten Wert ersetzen (Zeile oder Kolonne)
 - 4) Daten markieren (beste Lösung)

Umgang mit Ausreissern

Problem und Lösung, "non pivoted data" (1)

- Problem: Was machen wir, wenn Ausreisser vorkommen?
Definition: Siehe [Wikipedia](#)
- Pandas Lösungsansätze
 - 1) Sie zuerst mit describe() entdecken
 - 2) Daten visualisieren
 - 3) Ausreisser eventuell löschen...
 - 1) Von Hand
 - 2) Per Programm, wie [hier](#)

```
outlierdata.csv
"Landwirtschaftliche Betriebe und Beschäftigte nach Kanton"
"Variable";"Kanton";"Betriebssystem";"1990";"2000";"2010";"2020"
"Beschäftigte Total";"Zürich";"Konventionelle Betriebe";18832;13832;10907;9227
"Beschäftigte Total";"Bern / Berne";"Konventionelle Betriebe";52640;38082;123456789;27658
"Beschäftigte Total";"Luzern";"Konventionelle Betriebe";19821;15921;13638;12004
```

	Kanton	1990	2000	2010	2020
0	Zürich	18832	13832	10907	9227
1	Bern / Berne	52640	38082	123456789	27658
2	Luzern	19821	15921	13638	12004

Umgang mit Ausreissern

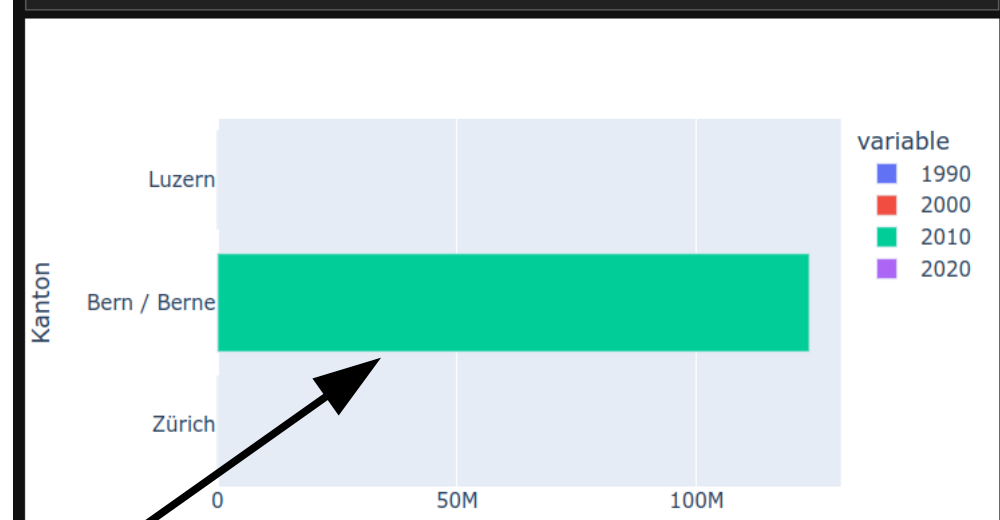
Problem und Lösung, "non pivoted data" (2)

- Jetzt sieht man deutlich, dass 2010 im Kanton Bern etwas mit den Daten schief gelaufen ist.
- Aber... was machen wir jetzt?
- Spannende Diskussion [hier](#)
- Keine gute Lösung. SBB Infra DLFW: Daten markieren

```
outlierdata_df.describe()
```

	1990	2000	2010	2020
count	3.000000	3.000000	3.000000e+00	3.000000
mean	30431.000000	22611.666667	4.116044e+07	16296.333333
std	19239.914007	13438.355194	7.127072e+07	9936.978129
min	18832.000000	13832.000000	1.090700e+04	9227.000000
25%	19326.500000	14876.500000	1.227250e+04	10615.500000
50%	19821.000000	15921.000000	1.363800e+04	12004.000000
75%	36230.500000	27001.500000	6.173521e+07	19831.000000
max	52640.000000	38082.000000	1.234568e+08	27658.000000

```
px.bar(outlierdata_df, x = ['1990', '2000', '2010', '2020'], y = 'Kanton')
```



Umgang mit Ausreissern

Problem und Lösung, "pivoted data" (1)

- describe() wird Ihnen nicht helfen...

	Kanton	Beschäftigte
Jahr		
1990	Zürich	18832
1990	Bern / Berne	52640
1990	Luzern	19821
2000	Zürich	13832
2000	Bern / Berne	38082
2000	Luzern	15921
2010	Zürich	10907
2010	Bern / Berne	123456789
2010	Luzern	13638
2020	Zürich	9227

```
outlierdata_pivoted_df.describe()
```

	Beschäftigte
count	1.200000e+01
mean	1.030745e+07
std	3.563280e+07
min	9.227000e+03
25%	1.322950e+04
50%	1.737650e+04
75%	3.026400e+04
max	1.234568e+08

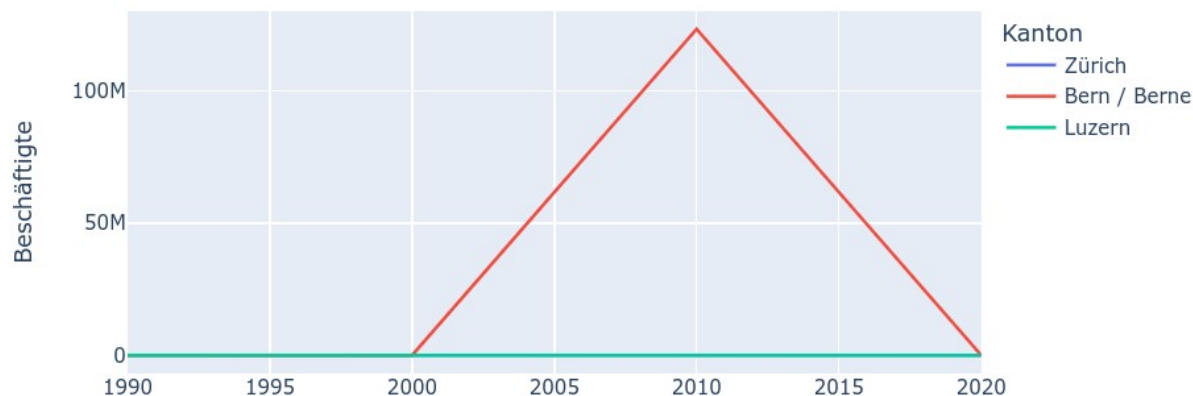
Umgang mit Ausreissern

Problem und Lösung, "pivoted data" (2)

- GroupBy + describe hilft
- Eine graphische Darstellung auch

```
outlierdata_pivoted_df.groupby('Kanton').describe()
```

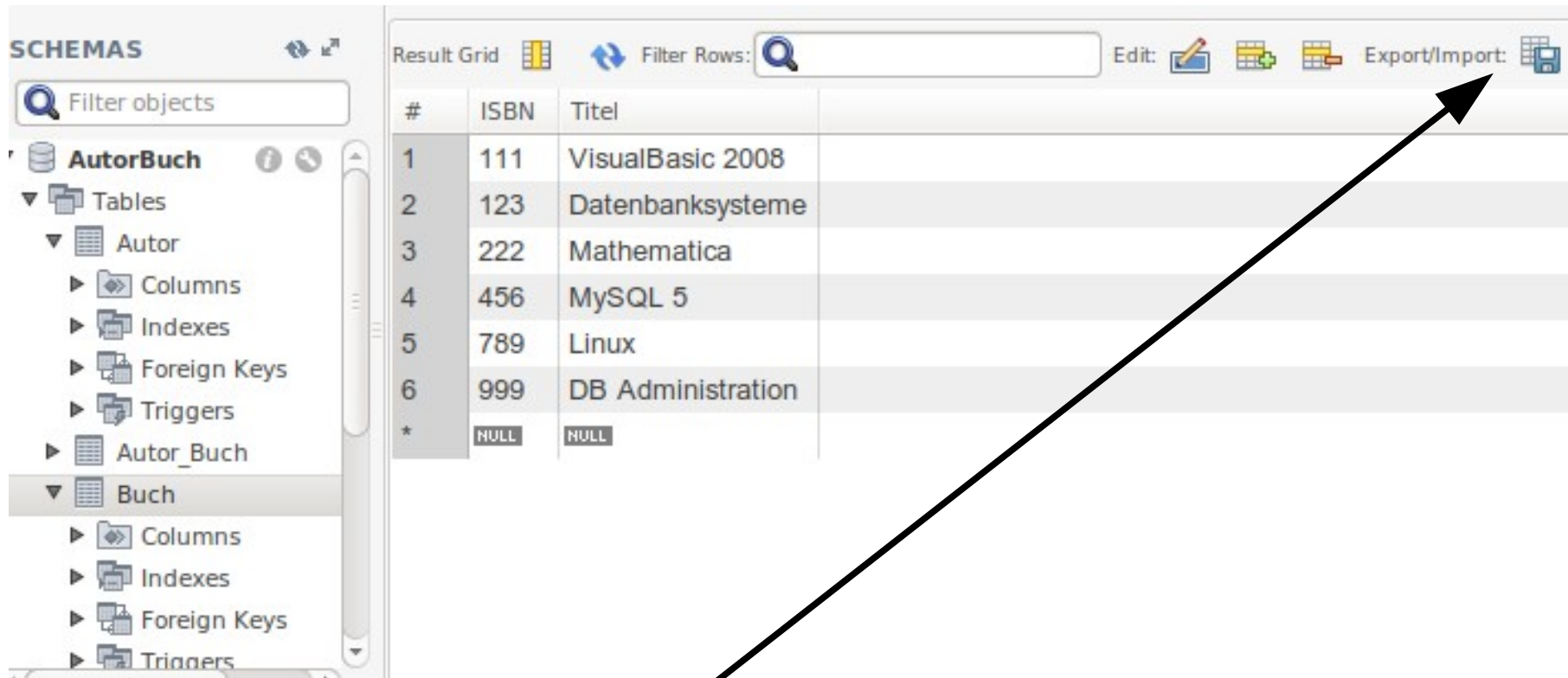
	count	mean	std	min	25%	50%	75%	max
Beschäftigte								
Kanton								
Bern / Berne	4.0	30893792.25	6.170867e+07	27658.0	35476.0	45361.0	30903677.25	123456789.0
Luzern	4.0	15346.00	3.388337e+03	12004.0	13229.5	14779.5	16896.00	19821.0
Zürich	4.0	13199.50	4.209569e+03	9227.0	10487.0	12369.5	15082.00	18832.0



Daten "joinen", mit dem AutorBuch Beispiel

Daten aus MySQL als CSV exportieren

- Wissen alle noch was ein Join ist?
- Am einfachsten Datenexport direkt mit MySQL Workbench



The screenshot shows the MySQL Workbench interface. On the left, the 'SCHEMAS' tree is expanded to show the 'AutorBuch' database, with the 'Buch' table selected. The main window displays a 'Result Grid' with the following data:

#	ISBN	Titel
1	111	VisualBasic 2008
2	123	Datenbanksysteme
3	222	Mathematica
4	456	MySQL 5
5	789	Linux
6	999	DB Administration
*	NULL	NULL

An arrow points from the bottom left towards the 'Export/Import' button in the top right corner of the result grid toolbar.

Daten "joinen", mit dem AutorBuch Beispiel Daten im Notebook direkt "joinen"

```
joined_autorBuch_df = pd.merge(autor_df, autorBuch_df, left_on='PersNr', right_on='PersonNr') \
    .drop('PersonNr', axis=1) \
    .merge(buch_df, on='ISBN')
joined_autorBuch_df
```

	PersNr	Vorname	Name	ISBN	Titel
0	12	Alfons	Kemper	123	Datenbanksysteme
1	34	Michael	Kofler	111	VisualBasic 2008
2	34	Michael	Kofler	222	Mathematica
3	34	Michael	Kofler	456	MySQL 5
4	34	Michael	Kofler	789	Linux

Übungen

Weitere Analysen und Vergleiche

- 1) Kopieren Sie
Notebook_ITP_DM0D8_Bio-LWBBetriebe-BEJUNE-V?.ipynb
nach
Notebook_ITP_DM0D8_Kon-LWBBetriebe-BEJUNE-V?.ipynb
und stellen Sie die Daten der konventionellen Betriebe dar.
Die Daten befinden sich hier
data/px-x-0702000000_107_KBETRIEB-Pivoted.csv
- 2) Erstellen Sie ein neues Notebook und bringen Sie das Beispiel der vorigen Seiten "*Daten joinen*", mit dem *AutorBuch Beispiel* zum Laufen.