Data Science 1

Hausarbeit

Prof. Dr. Christian Bockermann

Hochschule Bochum

Wintersemester 2023/2024

Prüfungsleistung - Hausarbeit

- Prüfungsleistung ist Hausarbeit
- Aufgabenstellung am 9.1.2024 um **9:00 Uhr**
- Bearbeitung in Jupyter Notebook
- Bearbeitungszeit bis 18.2.2024 um 23:59 Uhr
  (Abgabe: PDF in Moodle-Aufgabe hochladen)

- Gruppenarbeit mit max. 3 Personen möglich
- Notebook muss Namen + Matrikel-Nummern von allen Personen der Gruppe enthalten!
- Selbstorganisation der Gruppen, bitte bis 10.1. verbindlich mitteilen

Markdown

https://www.markdownguide.org/

Warum Visualisierung?

Warum Visualisierung von Daten?

- Überblick bei Exploration
- **Story Telling** mit den Daten
- Präsentation von Erkenntnissen aus Exploration

Interpretation von Daten

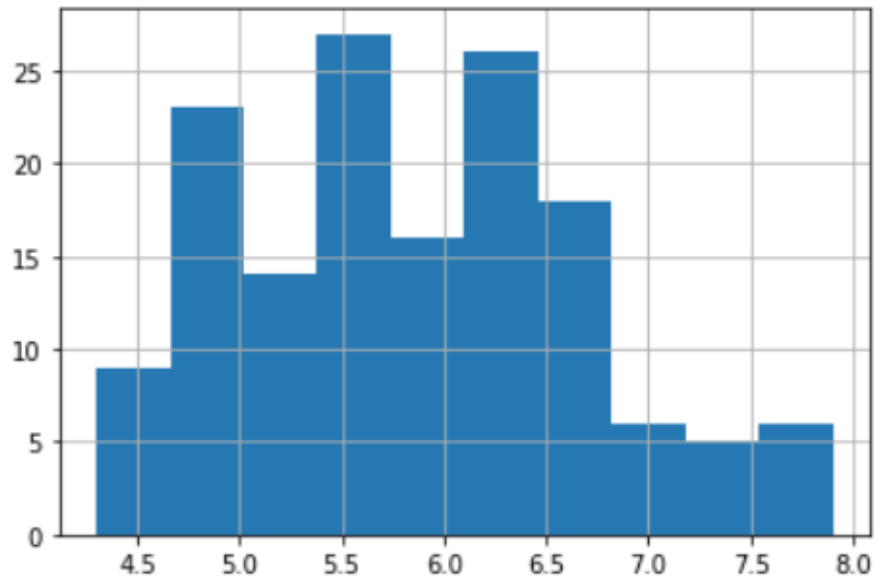
```
iris = pd.read_csv('iris.csv')
```

```
iris.describe()
```

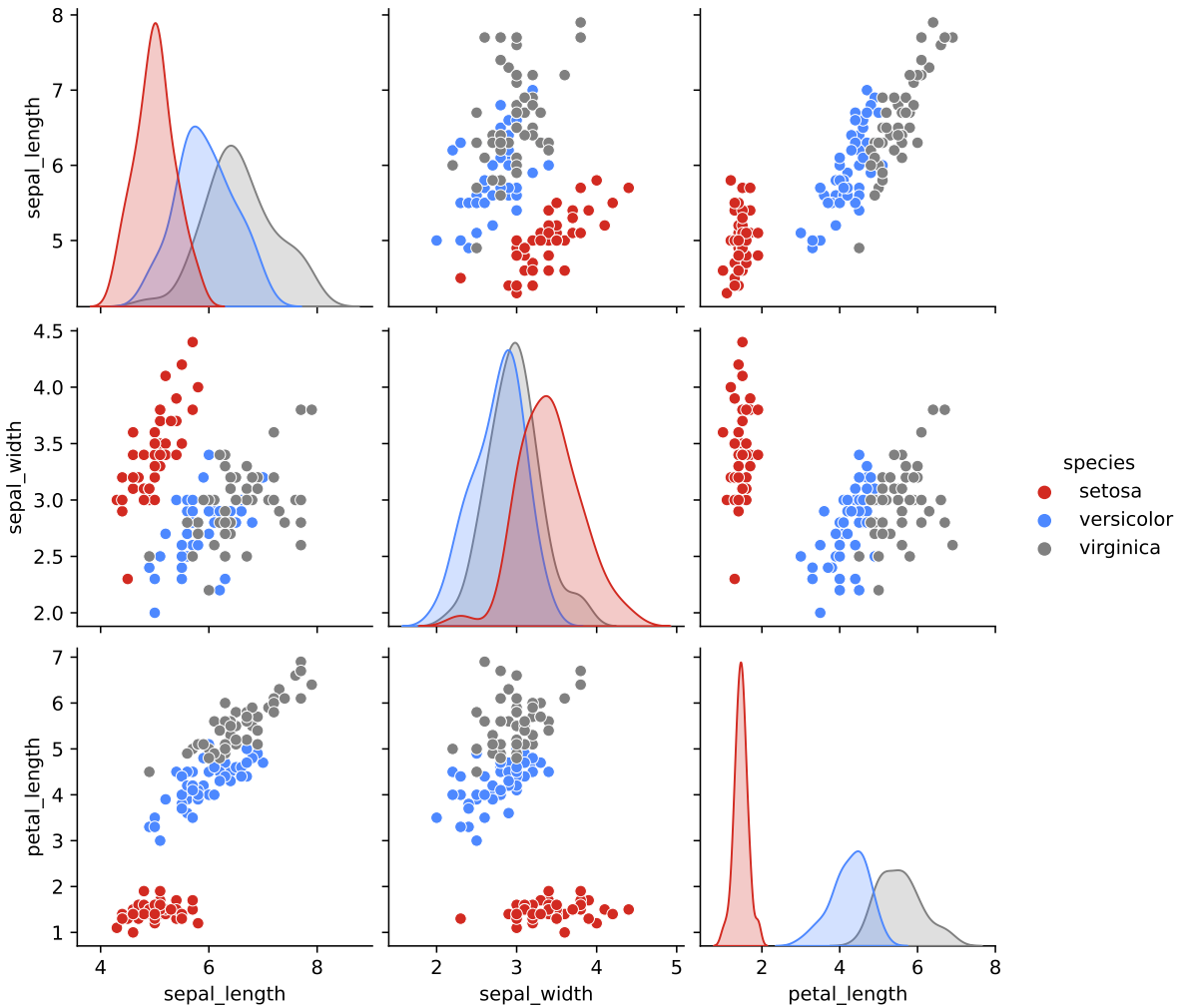
	sepal_length
count	150.000000
mean	5.843333
std	0.828066
min	4.300000
25%	5.100000
50%	5.800000
75%	6.400000
max	7.900000

```
iris['sepal_length'].hist()
```

```
<AxesSubplot:>
```



Komplexere Zusammenhänge darstellen



Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor?

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor? **Verteilung!**

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor? **Verteilung!**
- Wie “schwierig” ist der Datensatz?

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor? **Verteilung!**
- Wie “schwierig” ist der Datensatz? **Evaluierung!**

Exploration der Daten

- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor? **Verteilung!**
- Wie “schwierig” ist der Datensatz? **Evaluierung!**
- Welche Merkmale sind “gut” bzw. “hilfreich”?

Exploration der Daten

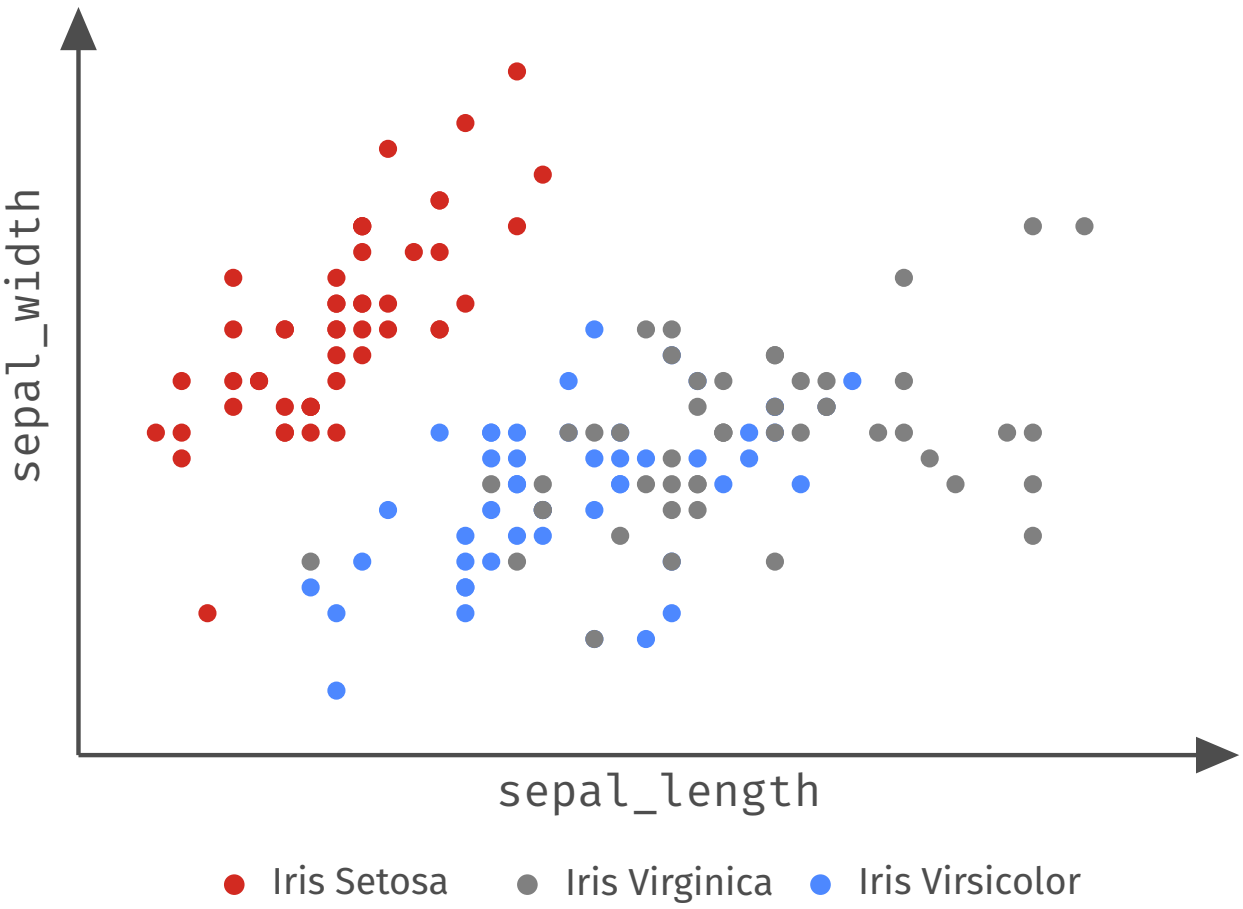
- Welche Merkmale (Spalten) hat der Datensatz?
- Welche Werte haben meine Merkmale? **Klassen!**
- Welche Werte kommen wie häufig vor? **Verteilung!**
- Wie “schwierig” ist der Datensatz? **Evaluierung!**
- Welche Merkmale sind “gut” bzw. “hilfreich”? **Optimierung!**

Beispiel: Iris Datensatz

sepal_length	sepal_width	petal_length	petal_width	species
6.3	2.3	4.4	1.3	versicolor
6.4	2.7	5.3	1.9	virginica
5.4	3.7	1.5	0.2	setosa
6.1	3.0	4.6	1.4	versicolor
5.0	3.3	1.4	0.2	setosa
5.0	2.0	3.5	1.0	versicolor

Iris Datensatz, [Fisher, 1988]

Beispiel: Iris Datensatz



Iris Datensatz - Wieviele Klassen (Arten)?

```
iris = pd.read_csv('iris.csv')
set(iris['species'])
```

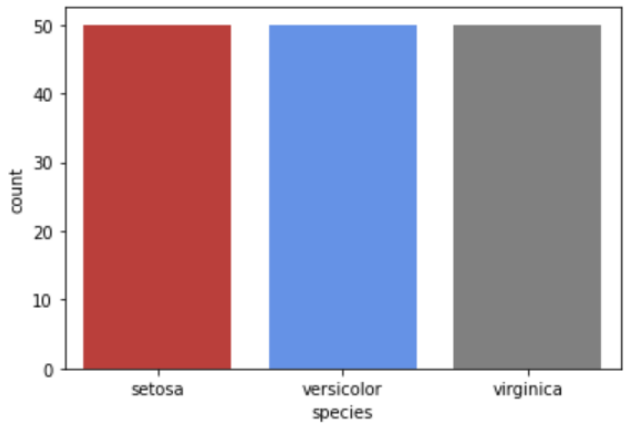
Ergebnis: {'setosa', 'versicolor', 'virginica'}

Iris Datensatz - Wieviele Klassen (Arten)?

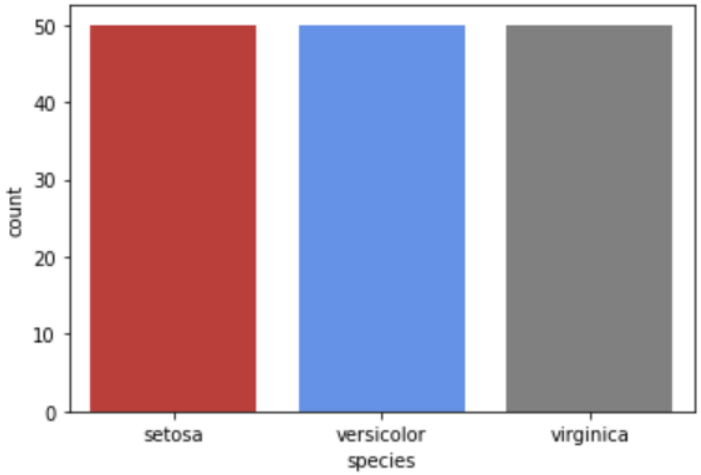
```
iris = pd.read_csv('iris.csv')
set(iris['species'])
```

Ergebnis: `{'setosa', 'versicolor', 'virginica'}`

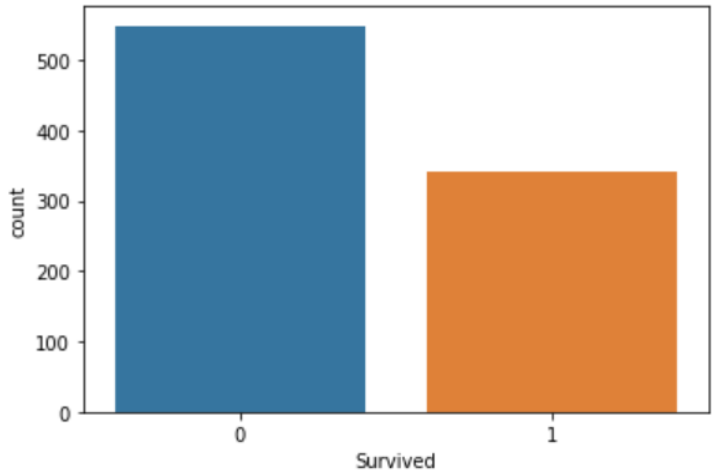
Als Grafik mit zusätzlichen Informationen:



Wie “schwierig” ist die Klassifikation?

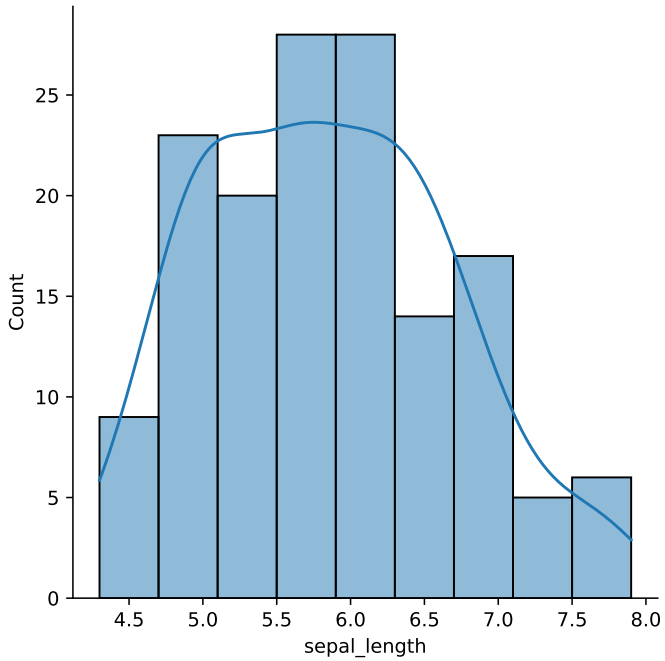


Iris Datensatz



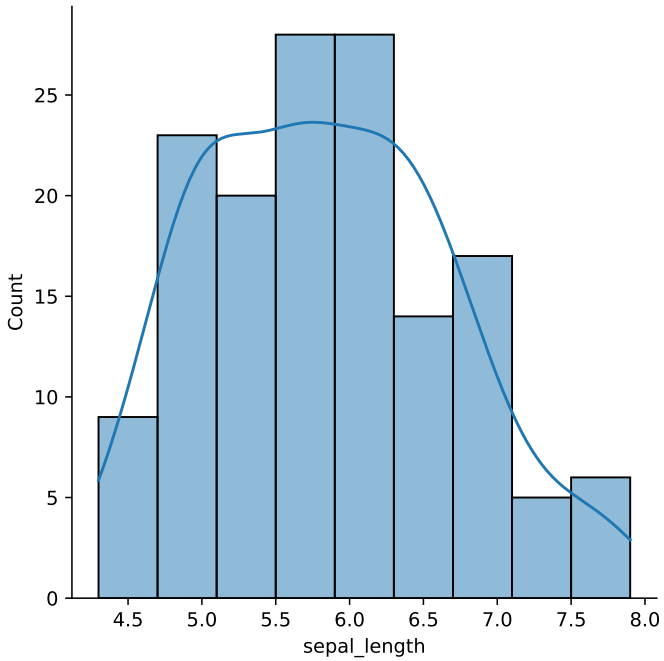
Titanic Datensatz

Wie ist die Verteilung eines bestimmten Merkmals?

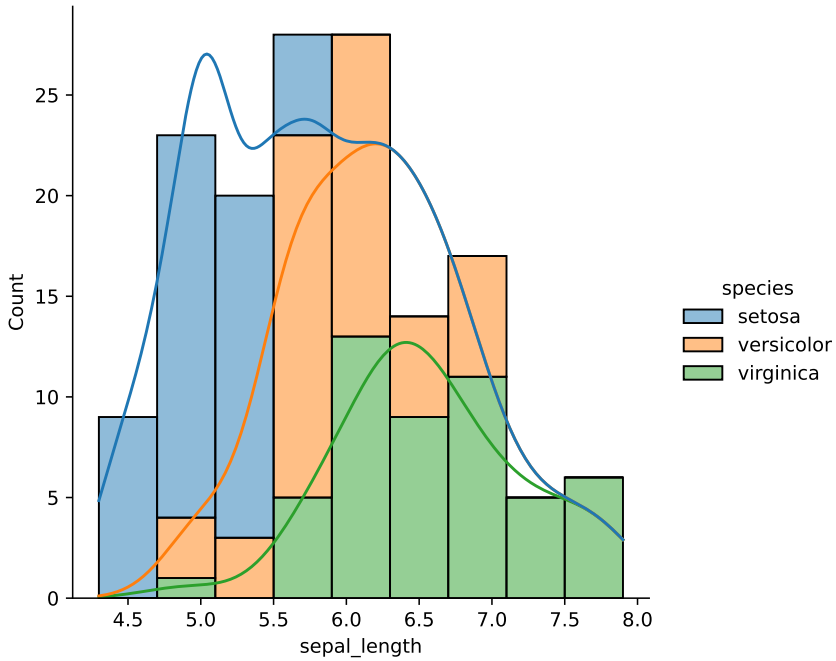


Gesamter Datensatz

Wie ist die Verteilung eines bestimmten Merkmals?

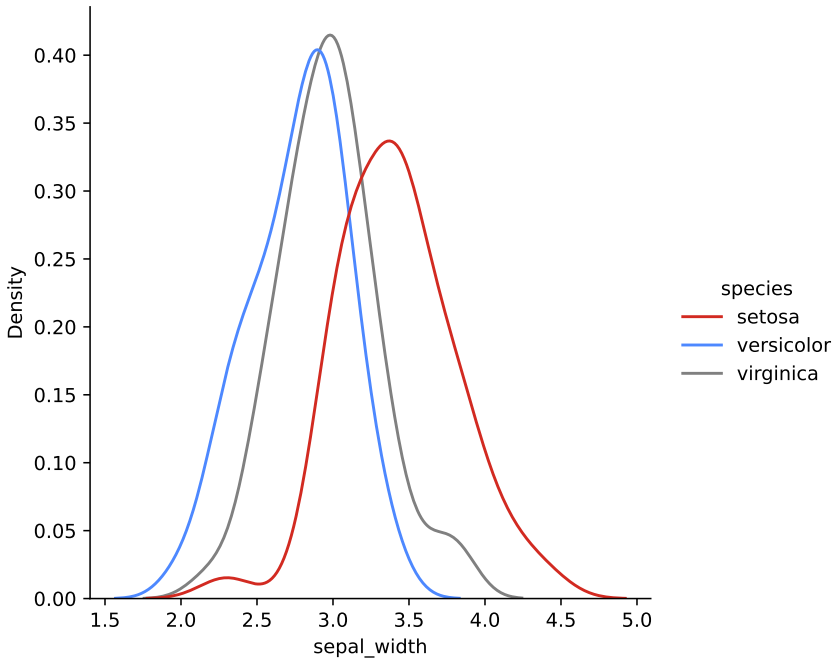


Gesamter Datensatz

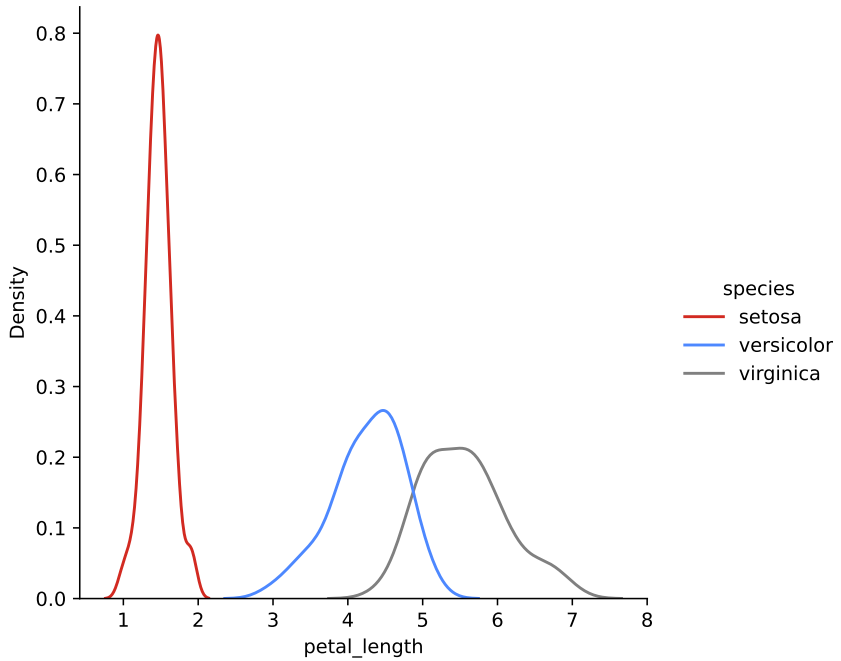


Nach Klassen

Was ist ggf. das wichtigere Merkmal?



`sepal_width`



`petal_length`

Visualisierung mit Seaborn

Das Seaborn Modul

- seaborn: statistical data visualization
- Entwickelt von Michael L. Waskom
- Modul für Datenvisualisierung
- Enthält zahlreiche Plot-Funktionen
- Kompatibel mit DataFrames (Pandas)

https://seaborn.pydata.org

```
import seaborn as sns
```

Verfügbare Seaborn Plots (Auszug)

- Histogramme (mit `sns.histplot(..)`)
- Verteilungen (mit `sns.displot(..)`)
- Bivariate Verteilungen (z.B. mit `sns.displot(..)`)
- uvm.

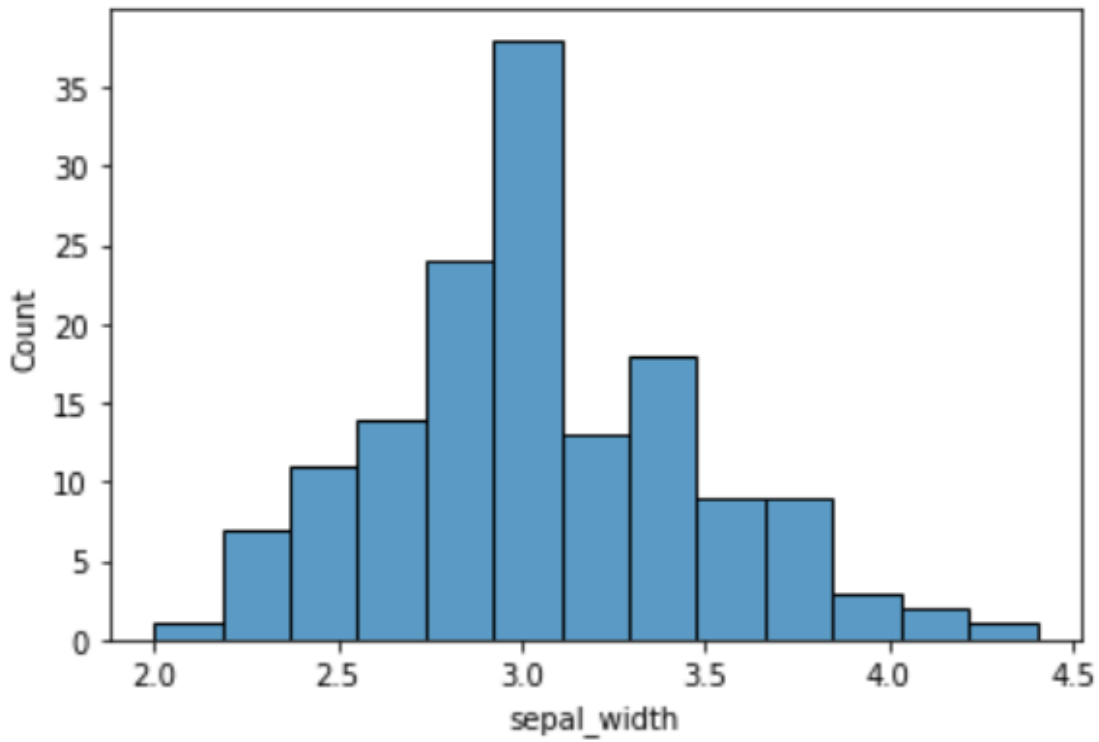
Grundsätzlicher Aufruf z.B. mit DataFrame:

```
# Parameter data muss immer da sein!
df = # dataframe!

sns.displot(data=df, x='spalte')
```

Histogram Plot

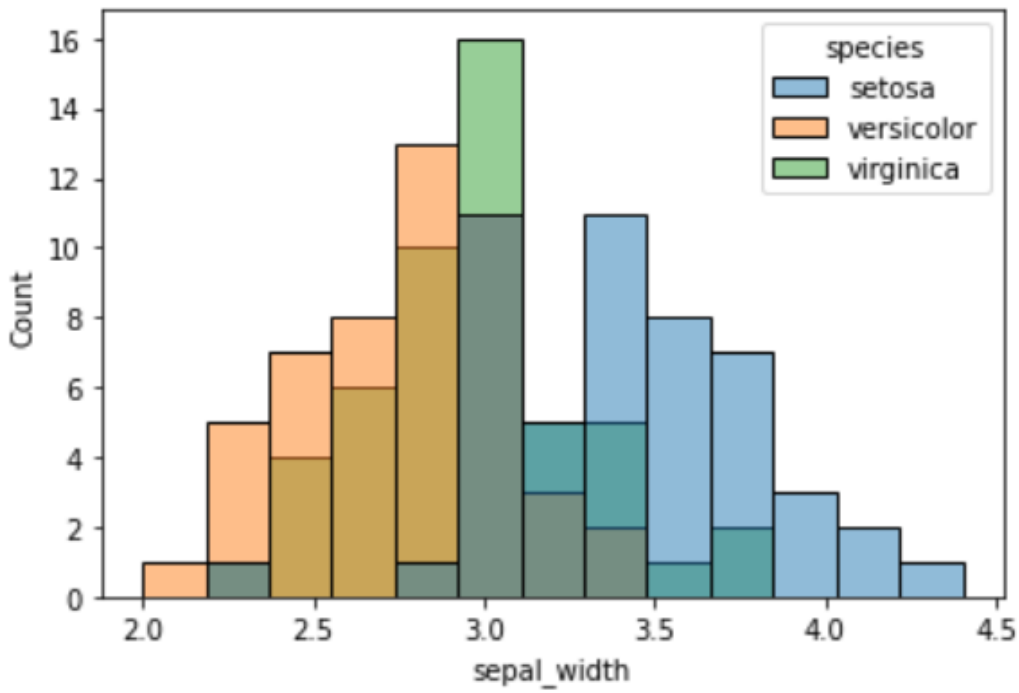
```
df = pd.read_csv('iris.csv')
sns.histplot(data=df, x='sepal_width')
```



Histogram Plot

Zusätzlicher Parameter hue für Unterteilung (z.B. nach Klassen)

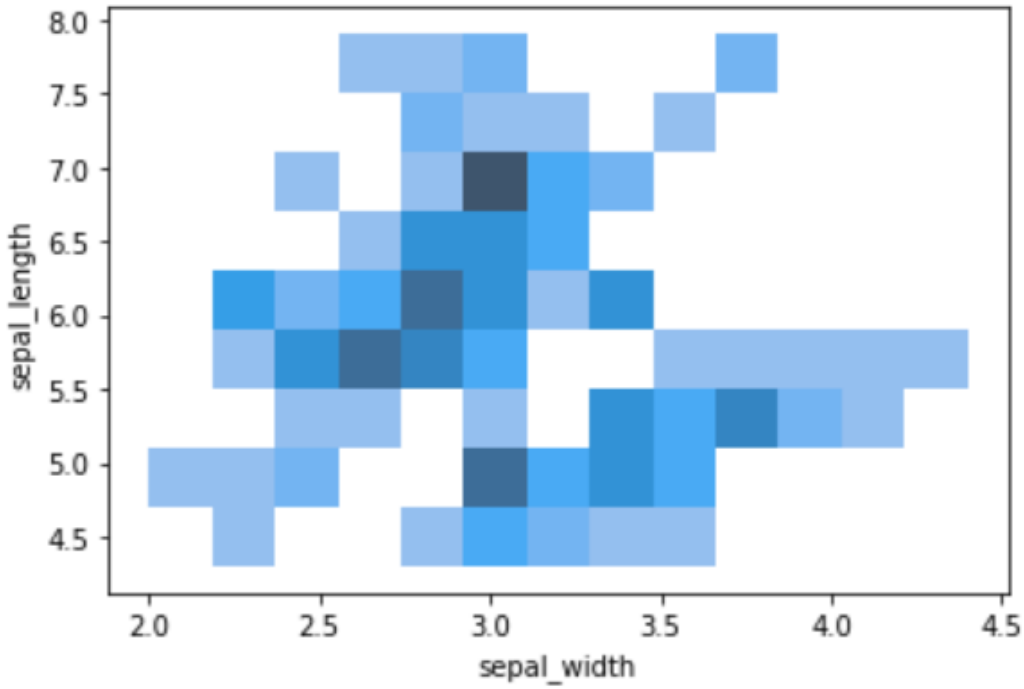
```
df = pd.read_csv('iris.csv')
sns.histplot(data=df, x='sepal_width', hue='species')
```



Histogram Plot

Histogram über zwei Merkmale:

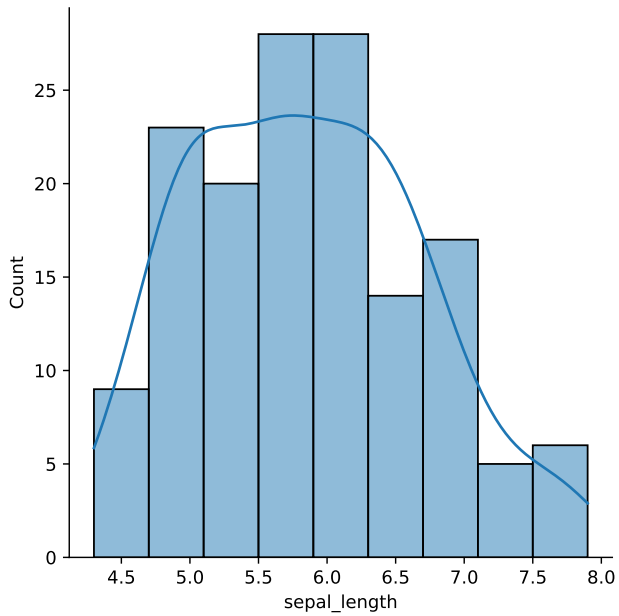
```
df = pd.read_csv('iris.csv')
sns.histplot(data=df, x='sepal_width', y='
                                sepal_length')
```



Verteilung von Merkmalen

Ähnlich wie Histogram, aber mit `sns.displot(..)`

```
sns.displot(data=df, x='sepal_length', kde=True)
```

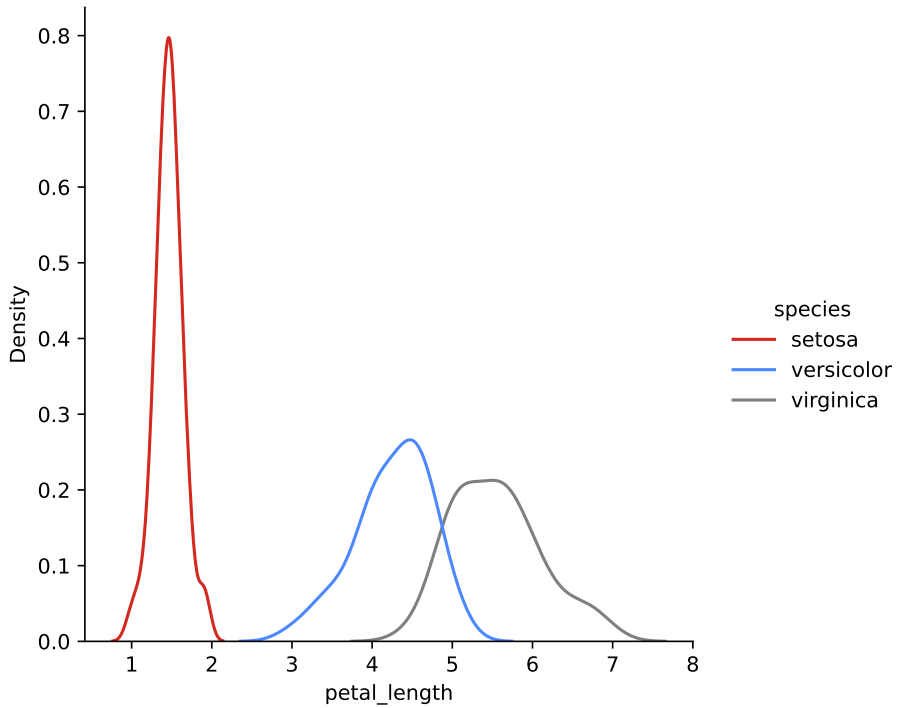


kde = *kernel density estimation* = Schätzung der Verteilungsfunktion

Verteilung von Merkmalen

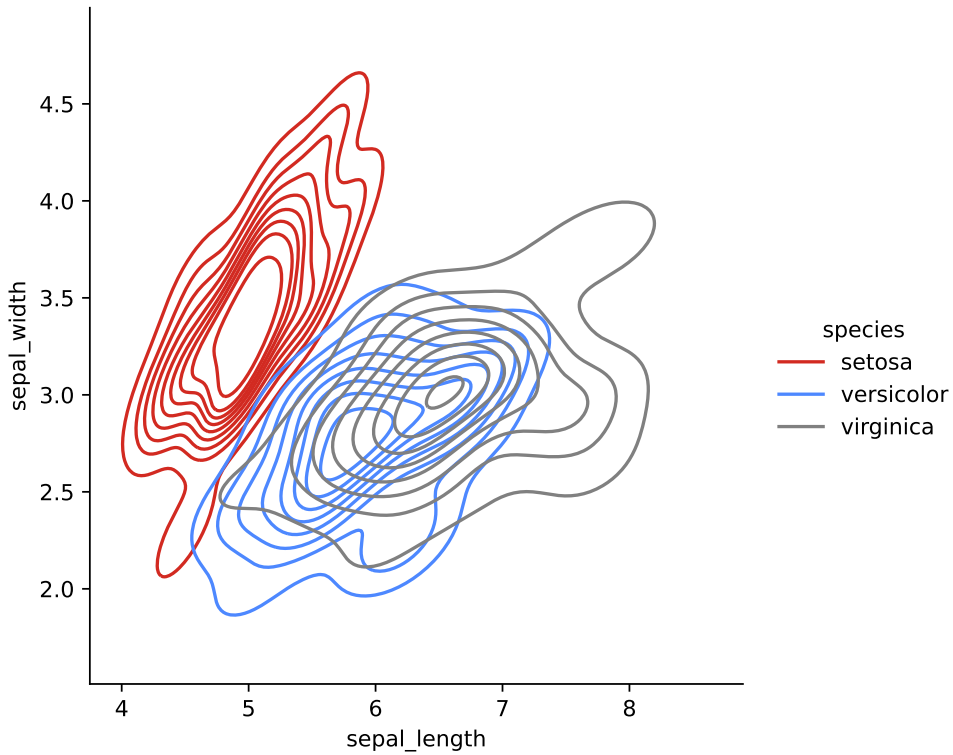
Parameter `kind='kde'` plottet nur die Verteilungsfunktion:

```
sns.displot(data=df, x='sepal_length', kind='kde',
                                hue='species')
```



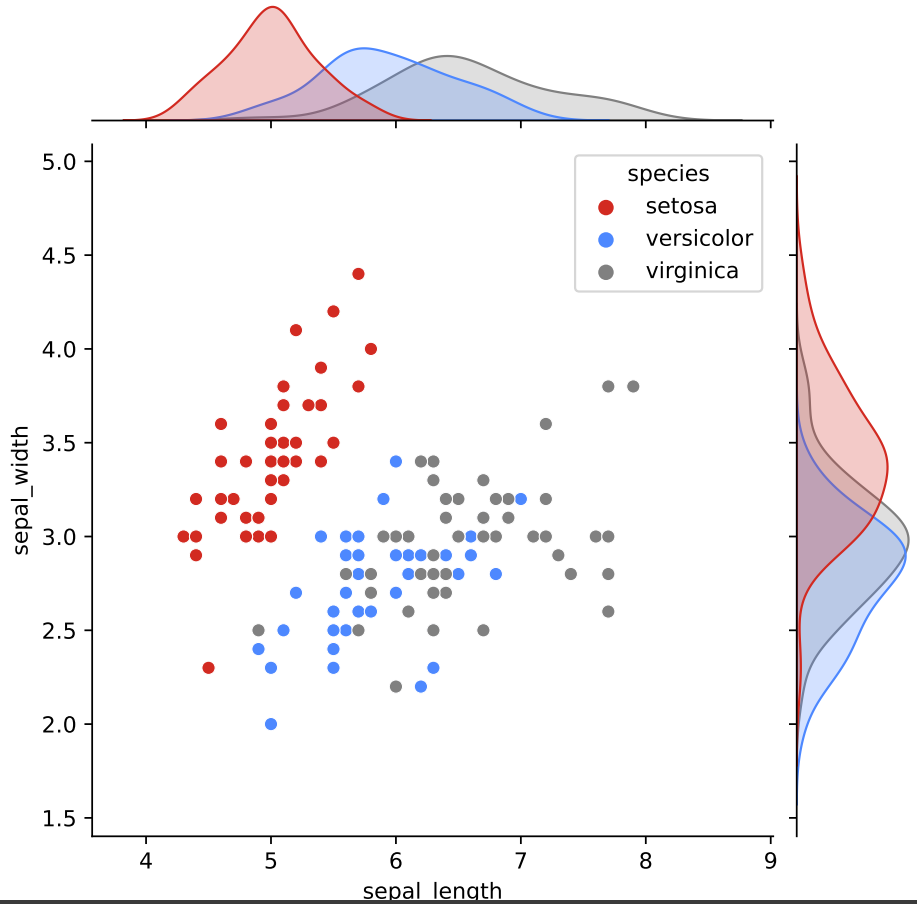
Verteilung von zwei Merkmalen

```
sns.displot(data=df,x='sepal_length',y='sepal_width',
            kind='kde',hue='species')
```



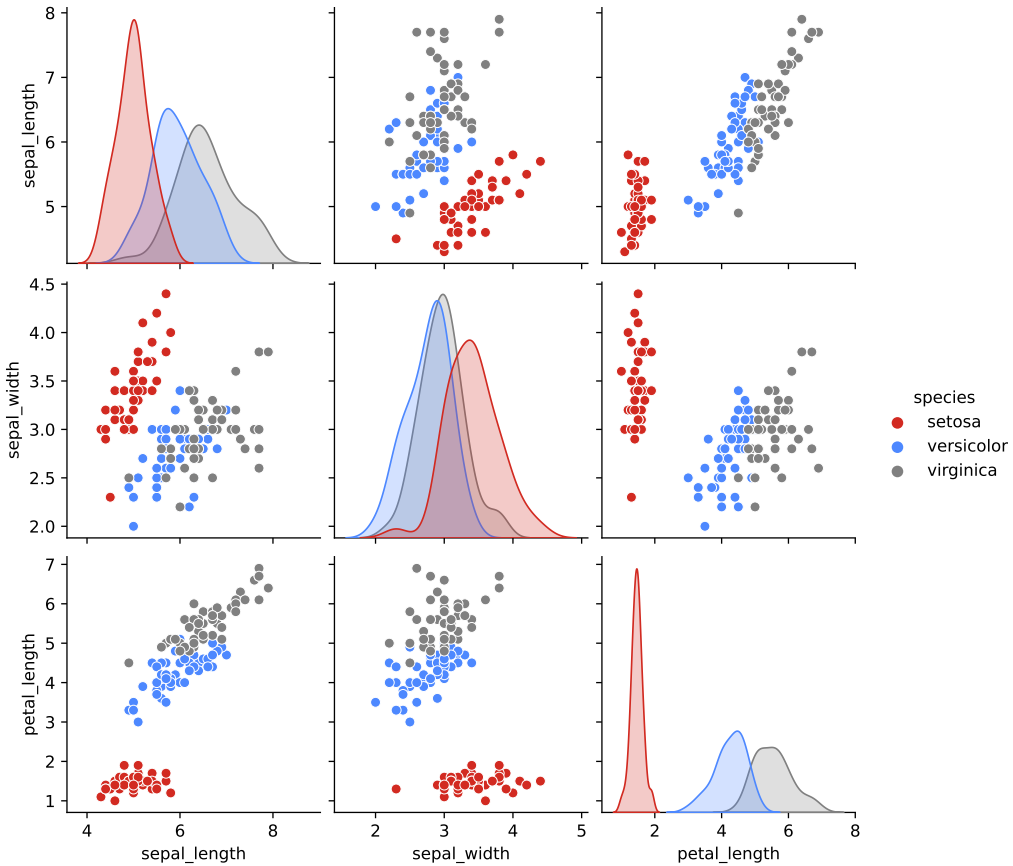
Verteilung von zwei Merkmalen mit jointplot

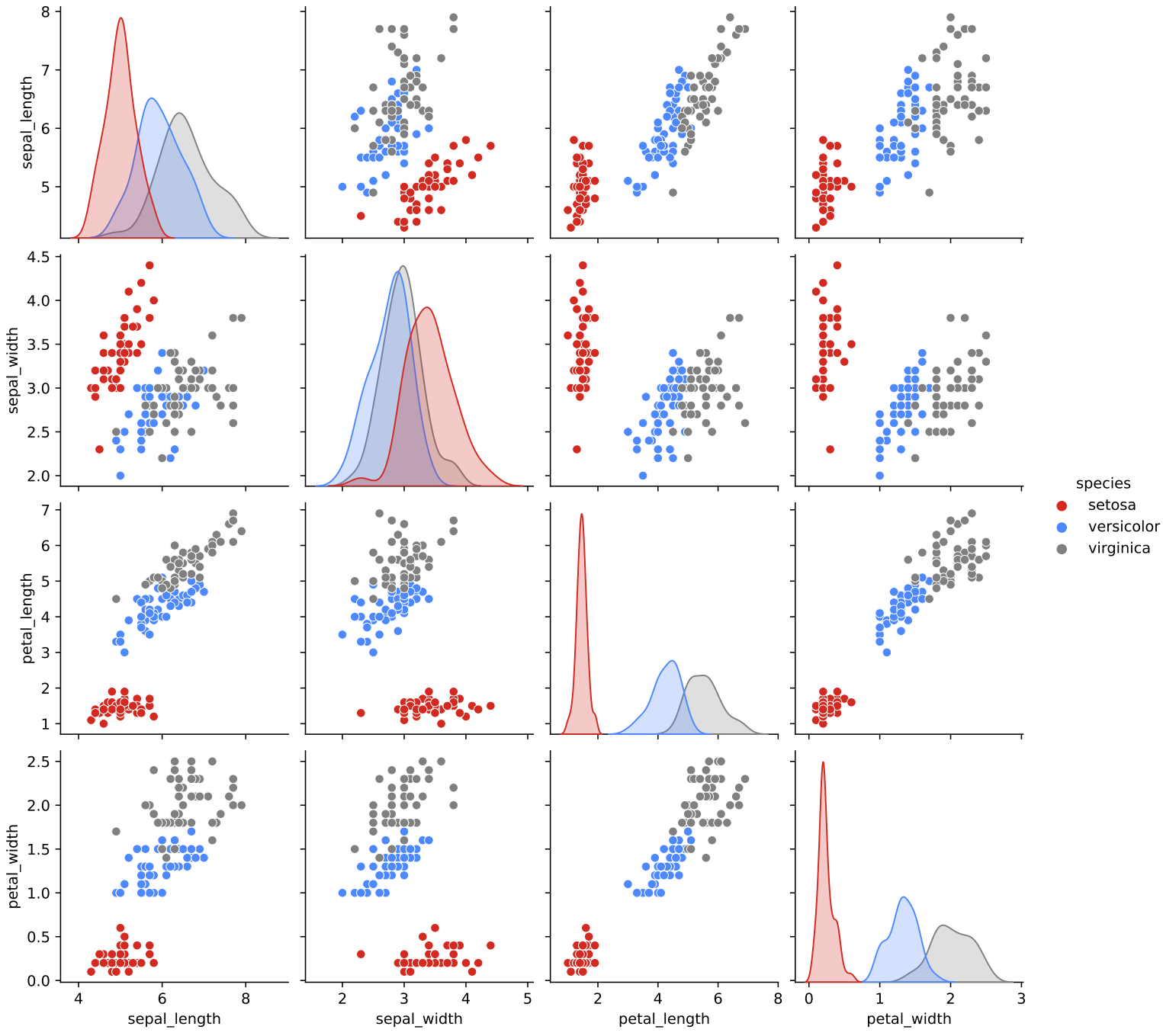
```
sns.jointplot(data=df,x='sepal_l',y='sepal_w',
              kind='kde',hue='species')
```



Verteilung von 2er-Kombinationen von Merkmalen (pairplot)

```
sns.pairplot(data=df,hue='species')
```




sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica

Viele weitere Visualisierungen

https://seaborn.pydata.org