DataCraft

4 Visualisierungen

DataCraft

Die relevanten Variablen eines Datensatzes werden häufig in einer Tabelle beschrieben

Für metrische Variablen werden häufig Anzahl, Spanne (Minimum und Maximum), Mittelwert mit Standardabweichung und der Median angegeben

Variable	n	Spanne	Mittelwert	Median
Schnabellänge	342 (2 NA)	32,1 - 59,6	43,9 ± 5,5	44,5
Schnabeltiefe	342 (2 NA)	13,1 - 21,5	17,2 ± 2,0	17,3
Flossenlänge	342 (2 NA)	172 - 231	200,9 ± 14,1	197
Körpergewicht	342 (2 NA)	2700 - 6300	4201,8 ± 802,0	4050

Python:

```
import pandas as pd
df.describe()
```

DataCraft

Die relevanten Variablen eines Datensatzes werden häufig in einer Tabelle beschrieben

Für nominale Variablen können nur die Häufigkeiten angegeben werden. Bei ordinalen Variablen hängt es von der Anzahl Ausprägungen ab, ob eine Auflistung der Werte und/oder Minimum, Median und Maximum (und ggf. Quartile) angegeben werden.

Variable	n
Art	344 (0 NA)
Adelie	152 (44,2%)
Gentoo	124 (36,0%)
Chinstrap	68 (19,8%)

Python:

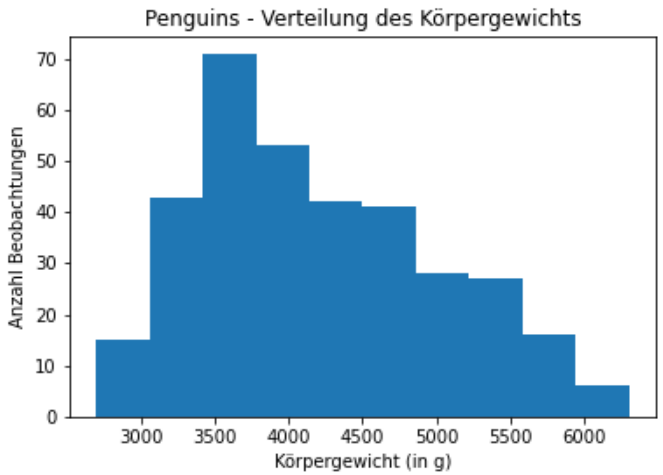
```
import pandas as pd
df["spalte"].value_counts()
```

DataCraft

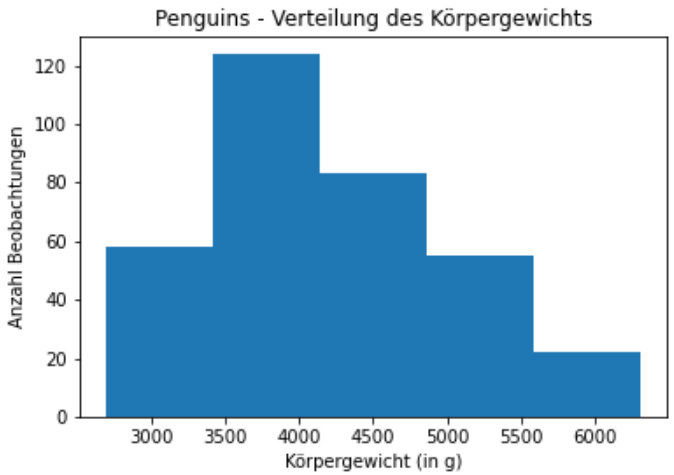
Histogramm = Häufigkeitsverteilung metrischer Merkmale

- Das Intervall wird in Klassen eingeteilt. Die Anzahl Klassen (gleich große Intervalle) oder deren Grenzen beeinflußen, wie die Grafik aussieht
- Dann wird gezählt, wie viele Werte in einer der Klassen liegt

10 Klassen



5 Klassen

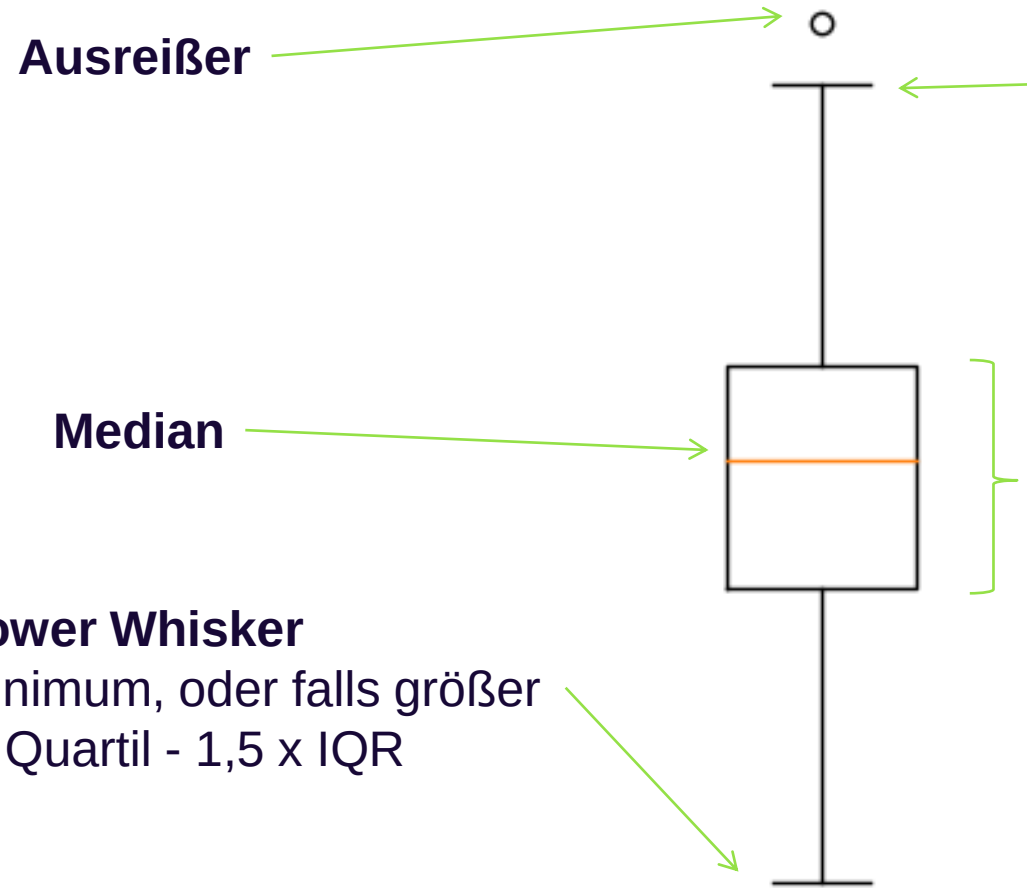


Pandas:

```
df["Spalte"].hist()
df["Spalte"].hist(bins=5,
                  grid=False)
```

DataCraft

Ein Boxplot visualisiert mehrere Kennzahlen für eine Variable



Ausreißer

Upper Whisker
Maximum, oder falls kleiner
3. Quartil + 1,5 x IQR

Median

Inter-Quartils-Range (IQR)
1. Quartil bis 3. Quartil
Hier liegen 50% der Daten

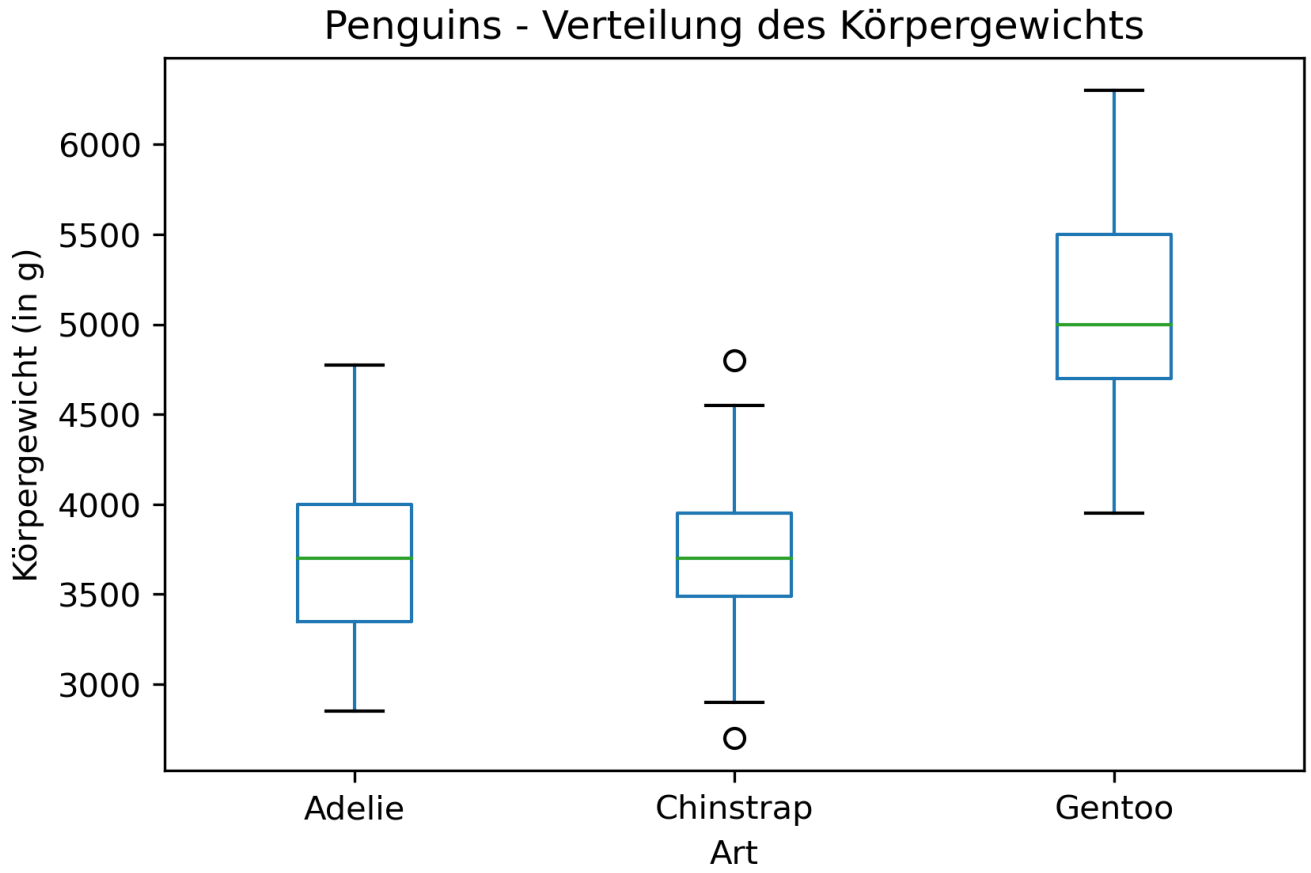
Lower Whisker
Minimum, oder falls größer
1. Quartil - 1,5 x IQR

Pandas:

```
df.boxplot("Spalte")
```

DataCraft

Boxplots sind gut geeignet, um die Verteilung einer Variable in mehreren Gruppen zu untersuchen



Pandas:

```
df.boxplot(column="Spalte",
           by="Gruppe")
```