EDA of the Iris Dataset

behrooz Filzadeh

2025-04-19

a) Beschreibung des Iris-Datensatzes

Der **Iris-Datensatz** ist ein bekannter Datensatz in der Statistik und im maschinellen Lernen, der erstmals 1935 vom britischen Biologen **Anderson** eingeführt wurde. Der Datensatz enthält Messungen von **150 Blumen** aus drei verschiedenen Arten der Gattung *Iris* (Setosa, Versicolor und Virginica). Jede Beobachtung umfasst vier kontinuierliche Merkmale: Kelchblattlänge, Kelchblattbreite, Blütenblattlänge und Blütenblattbreite.

Für mehr Informationen besuchen Sie die Wikipedia-Seite zum Iris-Datensatz (https://de.wikipedia.org/wiki/Iris-Blumendatensatz).

b) Laden des Iris-Datensatzes und explorative Datenanalyse für `Species == "setosa"`

```
# Laden der erforderlichen Bibliotheken
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

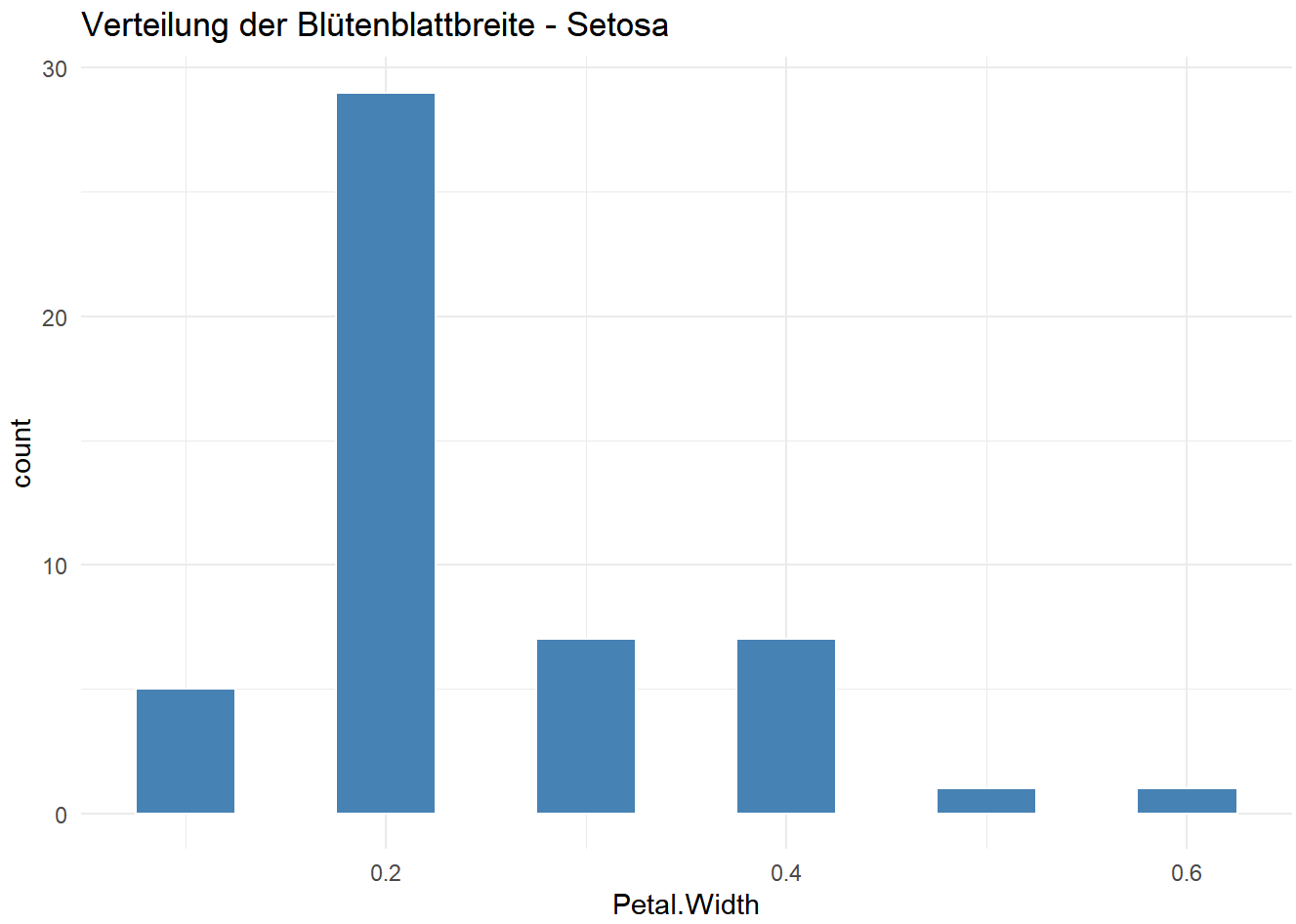
```r
library(tidyr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

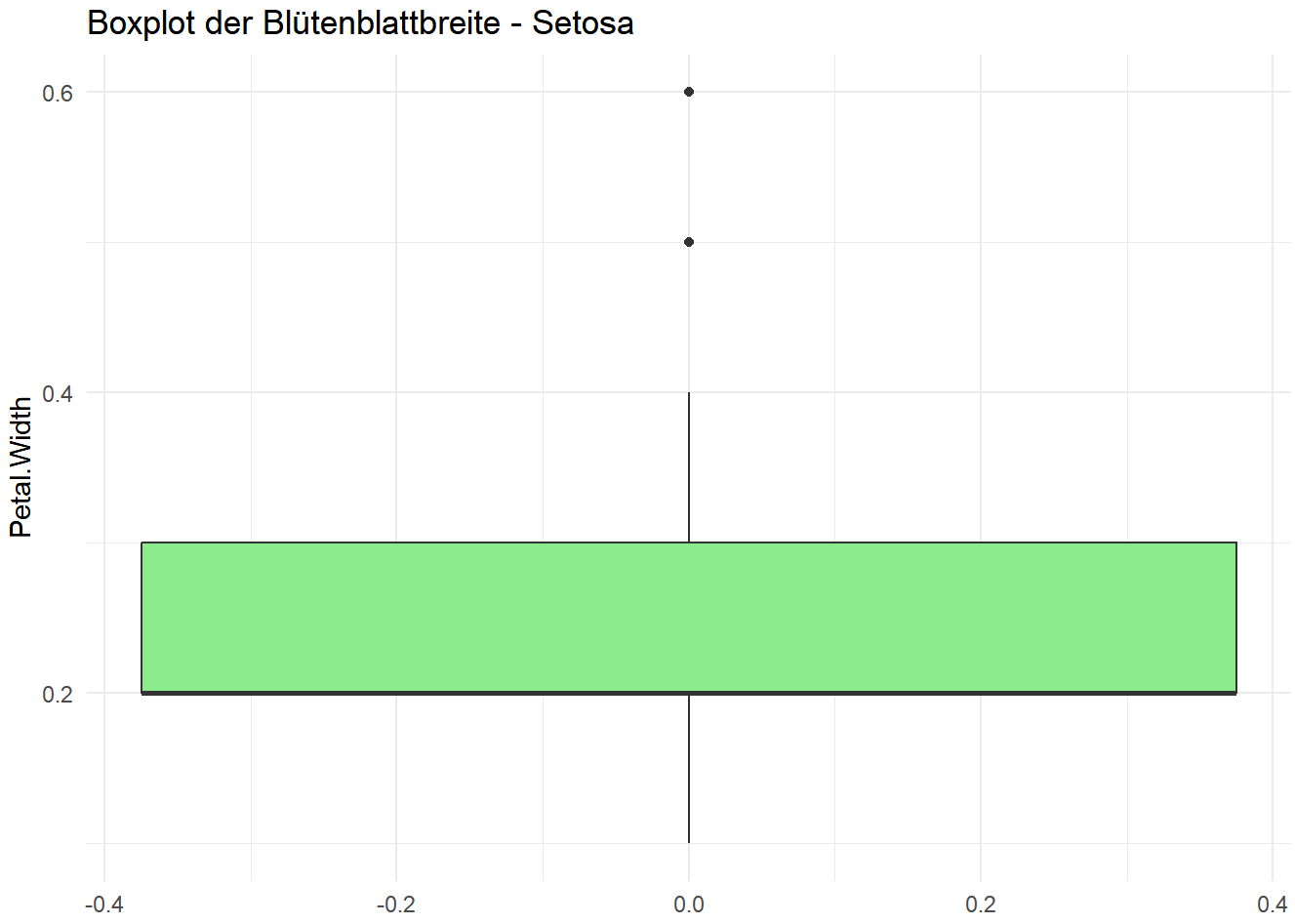
```r
# Laden des Iris-Datensatzes aus dem Datasets-Paket
data(iris)

# Auswahl der Daten für die Art "Setosa"
setosa <- iris %>% filter(Species == "setosa")

# Zusammenfassende Statistiken für Petal.Width in Setosa
summary(setosa$Petal.Width)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.200   0.200   0.246   0.300   0.600
```

```r
# Erstellen eines Histogramms für Petal.Width in der Art Setosa
ggplot(setosa, aes(x = Petal.Width)) +
  geom_histogram(binwidth = 0.05, fill = "steelblue", color = "white") +
  ggtitle("Verteilung der Blütenblattbreite - Setosa") +
  theme_minimal()
```



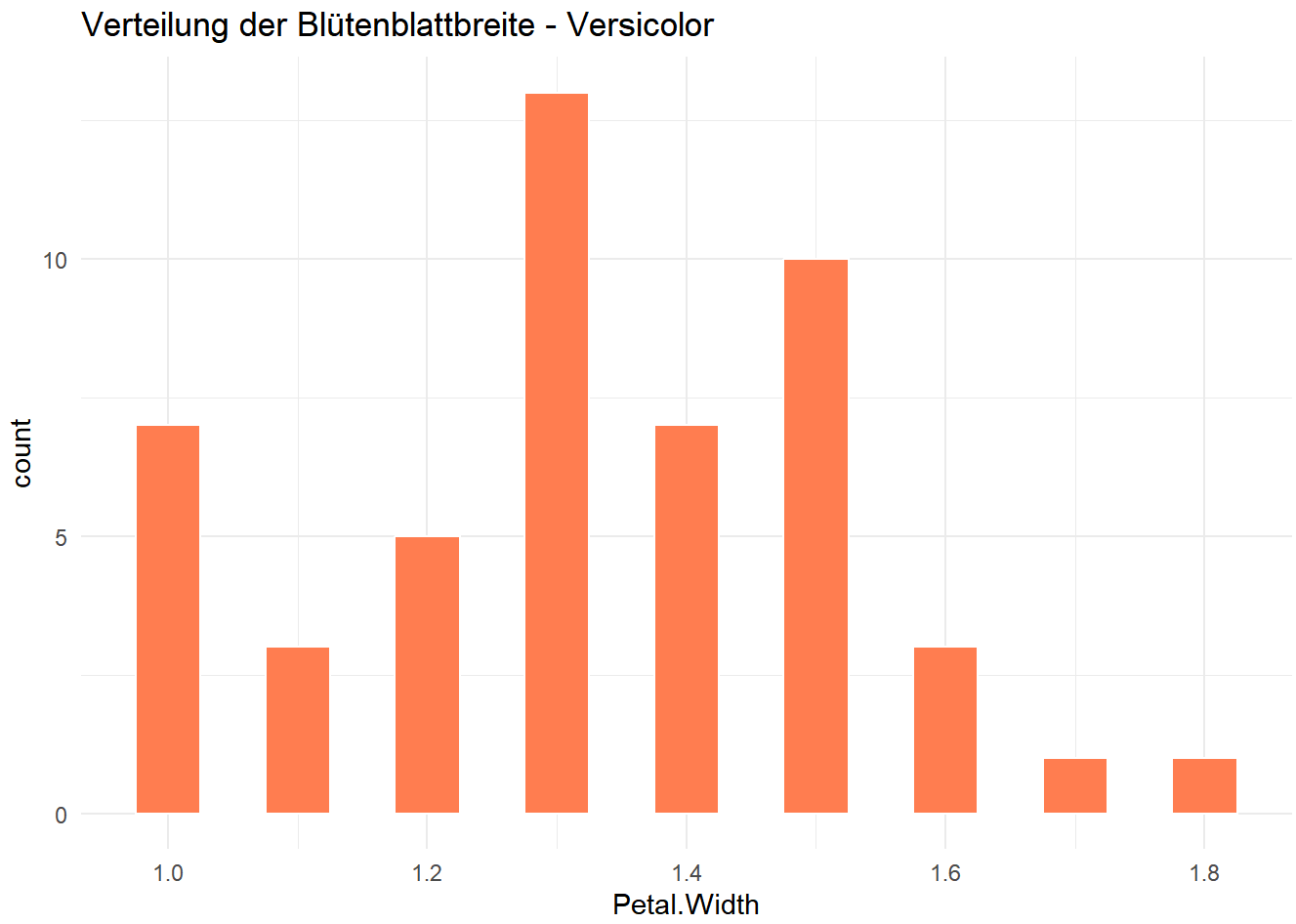

```
# Erstellen eines Boxplots für Petal.Width in der Art Setosa
ggplot(setosa, aes(y = Petal.Width)) +
  geom_boxplot(fill = "lightgreen") +
  ggtitle("Boxplot der Blütenblattbreite - Setosa") +
  theme_minimal()
```



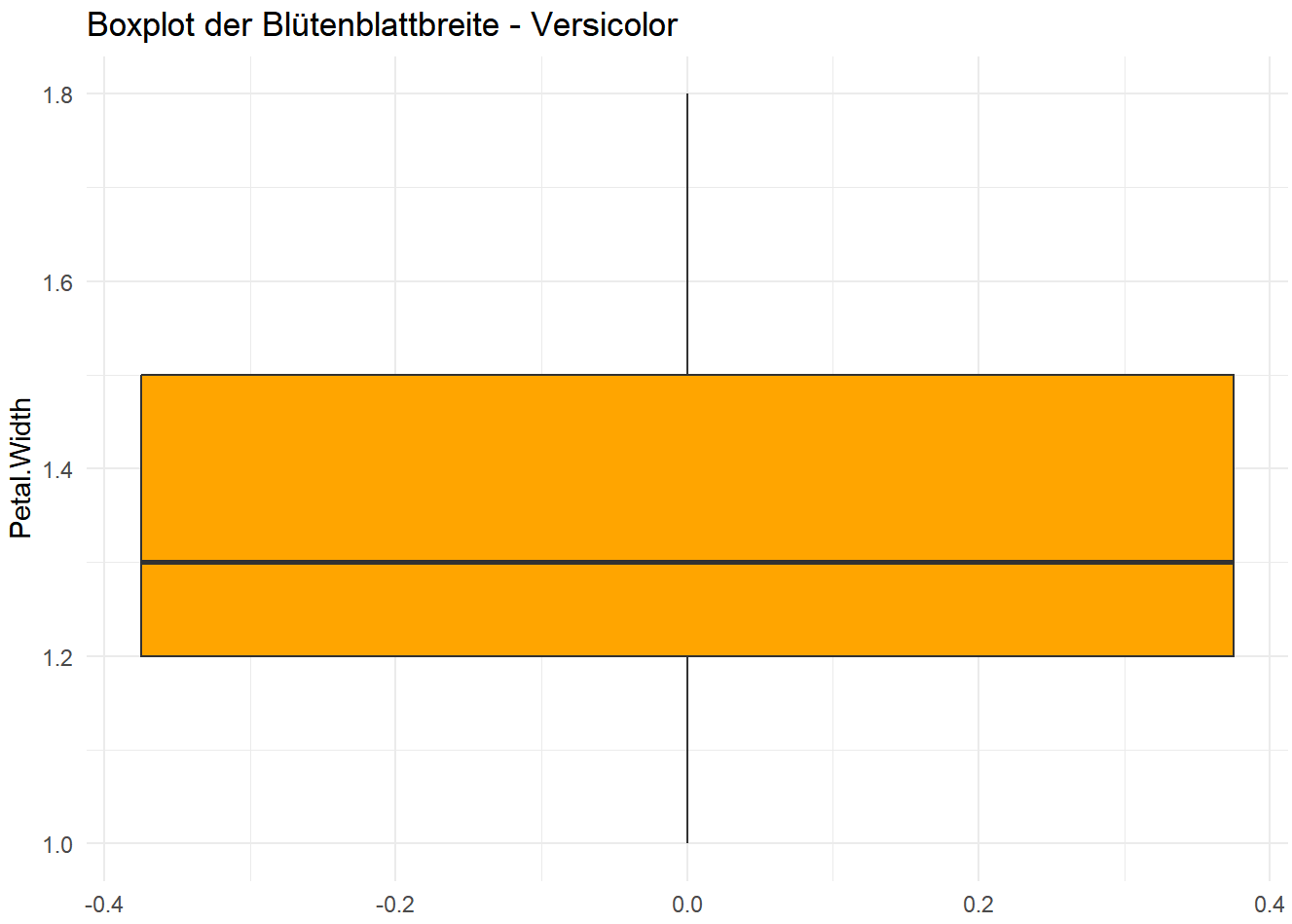

```
# Auswahl der Daten für die Art "Versicolor"
versicolor <- iris %>% filter(Species == "versicolor")

# Zusammenfassende Statistiken für Petal.Width in Versicolor
summary(versicolor$Petal.Width)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.200   1.300   1.326   1.500   1.800
```

```
# Erstellen eines Histogramms für Petal.Width in der Art Versicolor
ggplot(versicolor, aes(x = Petal.Width)) +
  geom_histogram(binwidth = 0.05, fill = "coral", color = "white") +
  ggtitle("Verteilung der Blütenblattbreite - Versicolor") +
  theme_minimal()
```



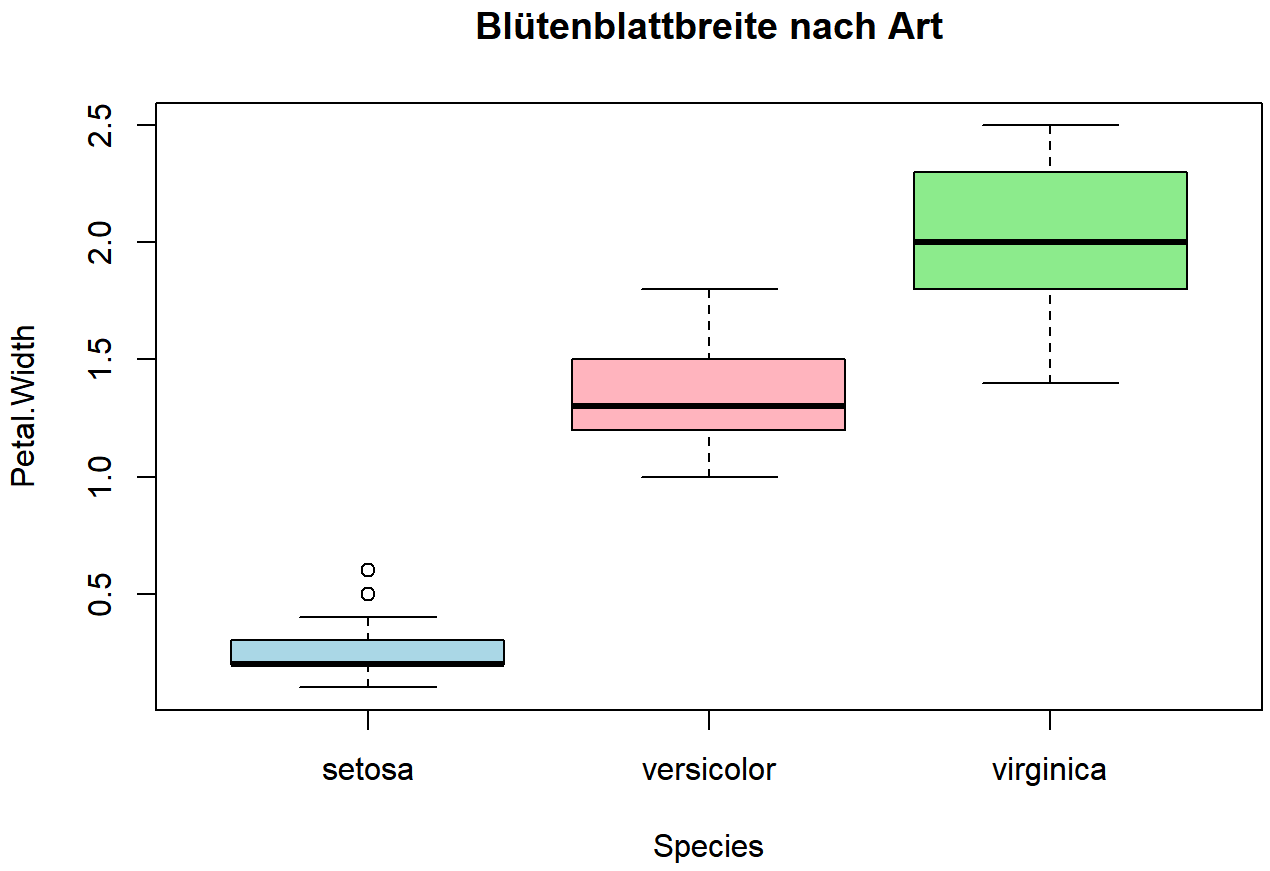

```
# Erstellen eines Boxplots für Petal.Width in der Art Versicolor
ggplot(versicolor, aes(y = Petal.Width)) +
  geom_boxplot(fill = "orange") +
  ggtitle("Boxplot der Blütenblattbreite - Versicolor") +
  theme_minimal()
```




```
# Verwendung der Funktion describeBy für deskriptive Statistiken nach Art
describeBy(iris[, 1:4], group = iris$Species)
```

```
##
##  Descriptive statistics by group
## group: setosa
##              vars  n mean   sd median trimmed  mad min max range skew kurtosis
## Sepal.Length    1 50 5.01 0.35    5.0    5.00 0.30 4.3 5.8   1.5 0.11    -0.45
## Sepal.Width     2 50 3.43 0.38    3.4    3.42 0.37 2.3 4.4   2.1 0.04     0.60
## Petal.Length    3 50 1.46 0.17    1.5    1.46 0.15 1.0 1.9   0.9 0.10     0.65
## Petal.Width     4 50 0.25 0.11    0.2    0.24 0.00 0.1 0.6   0.5 1.18     1.26
##                se
## Sepal.Length 0.05
## Sepal.Width  0.05
## Petal.Length 0.02
## Petal.Width  0.01
## ------------------------------------------------------------
## group: versicolor
##              vars  n mean   sd median trimmed  mad min max range  skew kurtosis
## Sepal.Length    1 50 5.94 0.52   5.90    5.94 0.52 4.9 7.0   2.1  0.10    -0.69
## Sepal.Width     2 50 2.77 0.31   2.80    2.78 0.30 2.0 3.4   1.4 -0.34    -0.55
## Petal.Length    3 50 4.26 0.47   4.35    4.29 0.52 3.0 5.1   2.1 -0.57    -0.19
## Petal.Width     4 50 1.33 0.20   1.30    1.32 0.22 1.0 1.8   0.8 -0.03    -0.59
##                se
## Sepal.Length 0.07
## Sepal.Width  0.04
## Petal.Length 0.07
## Petal.Width  0.03
## ------------------------------------------------------------
## group: virginica
##              vars  n mean   sd median trimmed  mad min max range  skew kurtosis
## Sepal.Length    1 50 6.59 0.64   6.50    6.57 0.59 4.9 7.9   3.0  0.11    -0.20
## Sepal.Width     2 50 2.97 0.32   3.00    2.96 0.30 2.2 3.8   1.6  0.34     0.38
## Petal.Length    3 50 5.55 0.55   5.55    5.51 0.67 4.5 6.9   2.4  0.52    -0.37
## Petal.Width     4 50 2.03 0.27   2.00    2.03 0.30 1.4 2.5   1.1 -0.12    -0.75
##                se
## Sepal.Length 0.09
## Sepal.Width  0.05
## Petal.Length 0.08
## Petal.Width  0.04
```

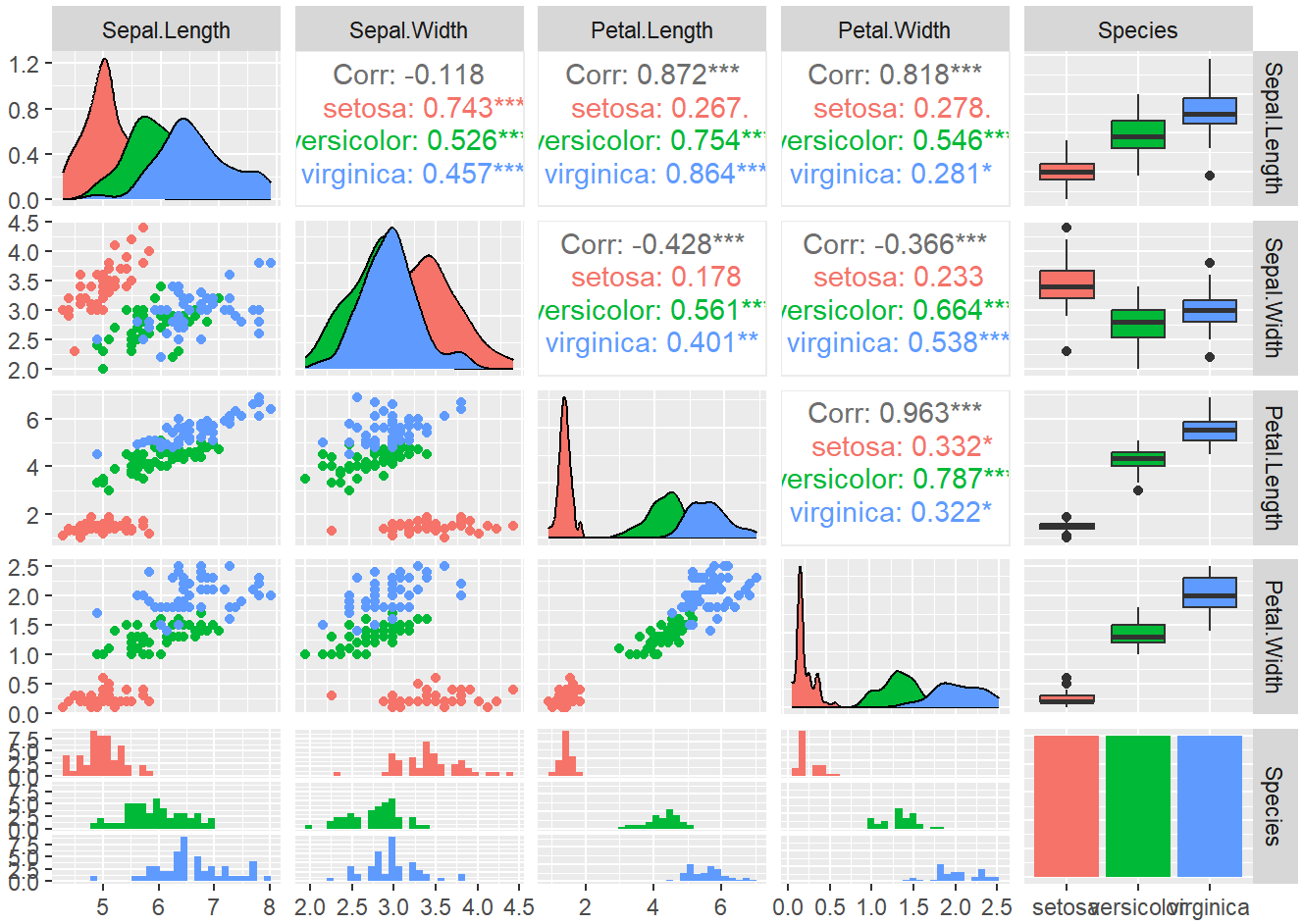
```
# Erstellen eines Boxplots für Petal.Width nach Art
boxplot(Petal.Width ~ Species, data = iris,
        main = "Blütenblattbreite nach Art",
        col = c("lightblue", "lightpink", "lightgreen"))
```



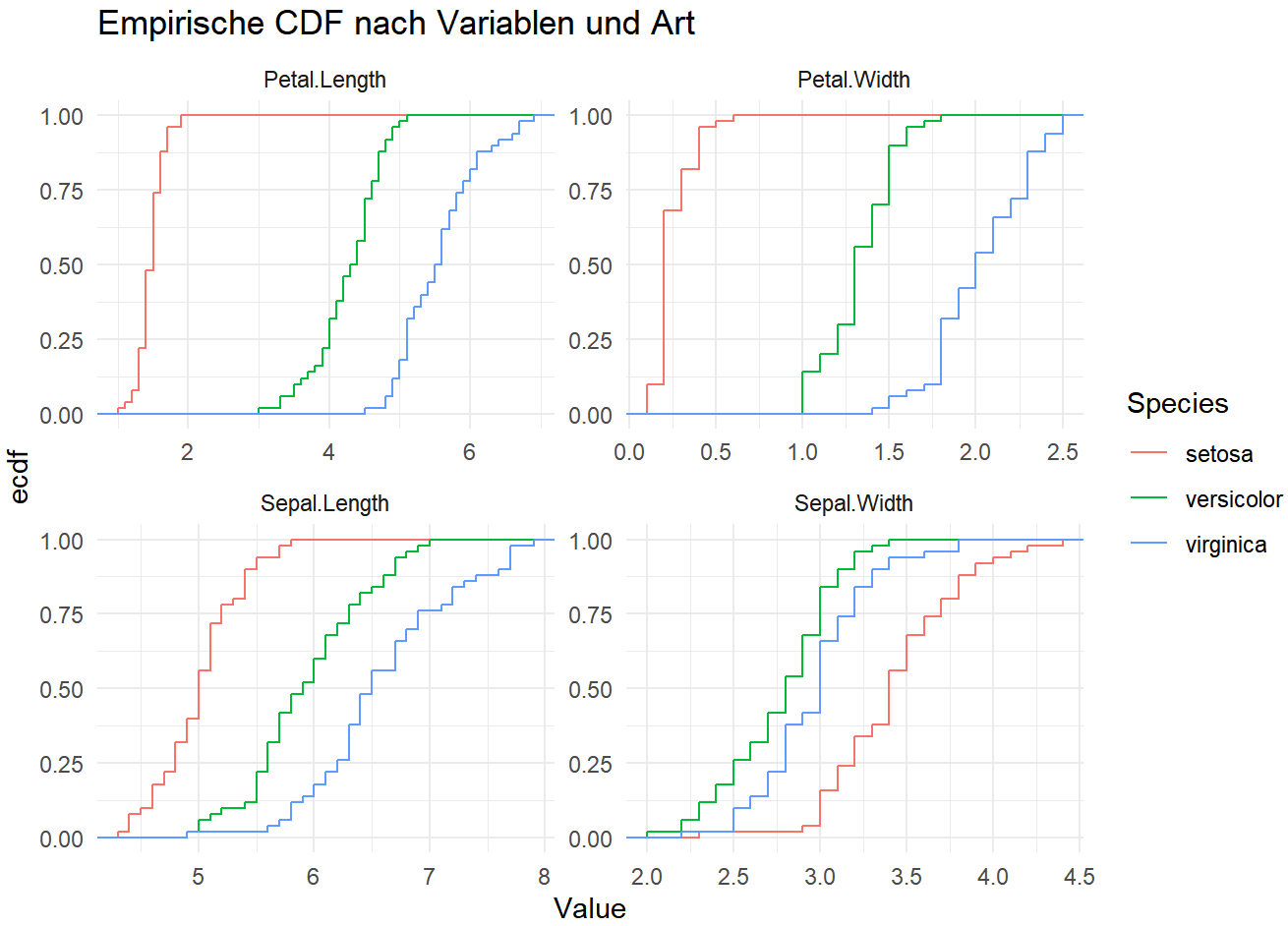

```
# Erstellen eines paarweisen Korrelationen-Diagramms mit ggpairs
ggpairs(iris, aes(color = Species))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
# Umformen der Daten in langes Format für die Erstellung des ECDF-Diagramms
iris_long <- iris %>%
  pivot_longer(cols = 1:4, names_to = "Variable", values_to = "Value")

# Erstellen eines ECDF-Diagramms nach Variablen und Art
ggplot(iris_long, aes(x = Value, color = Species)) +
  stat_ecdf() +
  facet_wrap(~ Variable, scales = "free") +
  labs(title = "Empirische CDF nach Variablen und Art") +
  theme_minimal()
```


Empirische CDF nach Variablen und Art
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
ecdf
Value
Species
setosa
versicolor
virginica
1.00
0.75
0.50
0.25
0.00
2
4
6
0.0
0.5
1.0
1.5
2.0
2.5
5
6
7
8
2.0
2.5
3.0
3.5
4.0
4.5