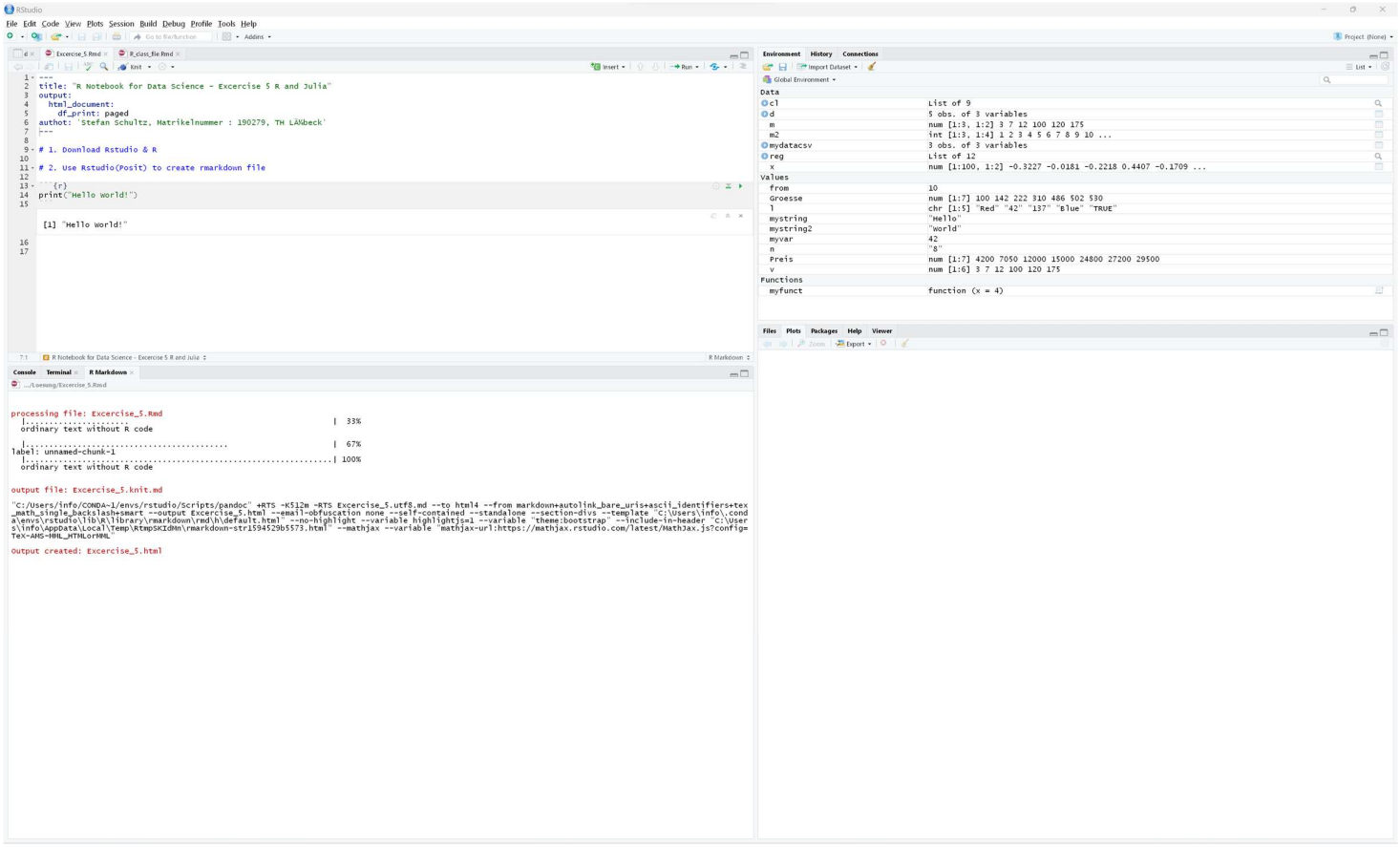
Stefan Schultz, Matrikelnr. 190279 stefan.schultz@stud.th-luebeck.de TH Lübeck

Aufgabe 1 + 2:

RStudio wurde heruntergeladen und auf einem Windows 11 Betriebssystem installiert.

Zum Test wurde eine Rmd-Notebook-Datei erstellt und mit einer simplen “Hello World!“ Ausgabe getestet. Siehe Screenshot:



Stefan Schultz, Matrikelnr. 190279 stefan.schultz@stud.th-luebeck.de TH Lübeck

Aufgabe 3:

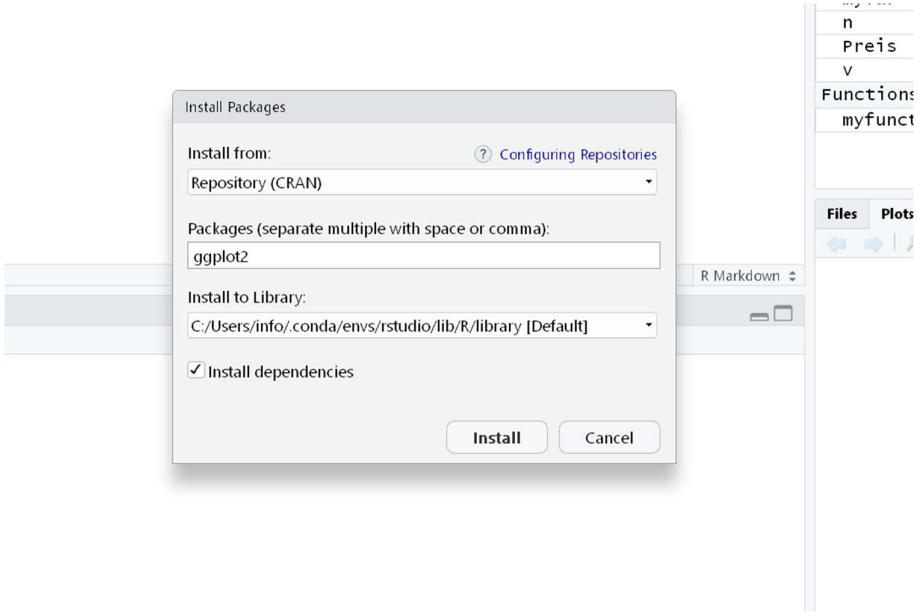
Alle Lösungen zu Aufgabe 3 sind alternativ in GitHub unter folgenden Repo zu finden:

https://github.com/stefanschultz/THL_DataScience

https://github.com/stefanschultz/THL_DataScience.git

- **download package ggplot2:**

  Installation des Packages kann über das Menü "Tools > Install Packages..." und Eingabe des Package-Namens mit Klick auf Install, sowie Angabe des Installationsverzeichnisses installiert werden.



```
    set_makevars                            html
    with_                                   html
    with_collate                            html
    with_connection                         html
    with_db_connection                      html
    with_dir                                html
    with_envvar                             html
    with_file                               html
    with_gctorture2                         html
    with_language                           html
    with_libpaths                           html
    with_locale                             html
    with_makevars                           html
    with_options                            html
    with_package                            html
Rd warning: C:/Users/info/AppData/Local/Temp/RtmpOiFzVa/R.INSTALL52c86da13e25/withr/man/with_package.Rd:68: file link '.libPaths' in package 'base' does n
ot exist and so has been treated as a topic
    with_par                                html
    with_path                               html
    with_rng_version                        html
    with_seed                               html
    with_sink                               html
    with_temp_libpaths                      html
    with_tempfile                           html
    with_timezone                           html
    withr                                   html
*** copying figures
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (withr)
ERROR: dependency 'lifecycle' is not available for package 'gtable'
* removing 'C:/Users/info/.conda/envs/rstudio/lib/R/library/gtable'
Warning in install.packages :
  installation of package ‘gtable’ had non-zero exit status
ERROR: dependency 'lifecycle' is not available for package 'scales'
* removing 'C:/Users/info/.conda/envs/rstudio/lib/R/library/scales'
Warning in install.packages :
  installation of package ‘scales’ had non-zero exit status
ERROR: dependencies 'gtable', 'lifecycle', 'MASS', 'scales' are not available for package 'ggplot2'
* removing 'C:/Users/info/.conda/envs/rstudio/lib/R/library/ggplot2'
Warning in install.packages :
  installation of package ‘ggplot2’ had non-zero exit status

The downloaded source packages are in
        ‘C:\Users\info\AppData\Local\Temp\RtmpuUHXw3\downloaded_packages’
> |
```

- **import the data("iris")**
  Herunterladen des Datensets von Iris „Iris.csv“ von kaggle und Speicherung auf lokaler Festplatte.

Laden der CSV-Datei mit der Library „readr“. Ausgabe des Klassennamens und der CSV-Spaltennamen:

```r
# use library readr for loading csv data from file.
library(readr)

# load Iris data set from csv file
Iris <- read.csv('./input/Iris.csv')

# output class name
print(class(Iris))

# print head of Iris class (column names)
head(Iris)
```

R Console

data.frame
6 x 6

	Id <int>	SepalLengthCm <dbl>	SepalWidthCm <dbl>	PetalLengthCm <dbl>	PetalWidthCm <dbl>	Species <fctr>
1	1	5.1	3.5	1.4	0.2	Iris-setosa
2	2	4.9	3.0	1.4	0.2	Iris-setosa
3	3	4.7	3.2	1.3	0.2	Iris-setosa
4	4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	5.0	3.6	1.4	0.2	Iris-setosa
6	6	5.4	3.9	1.7	0.4	Iris-setosa

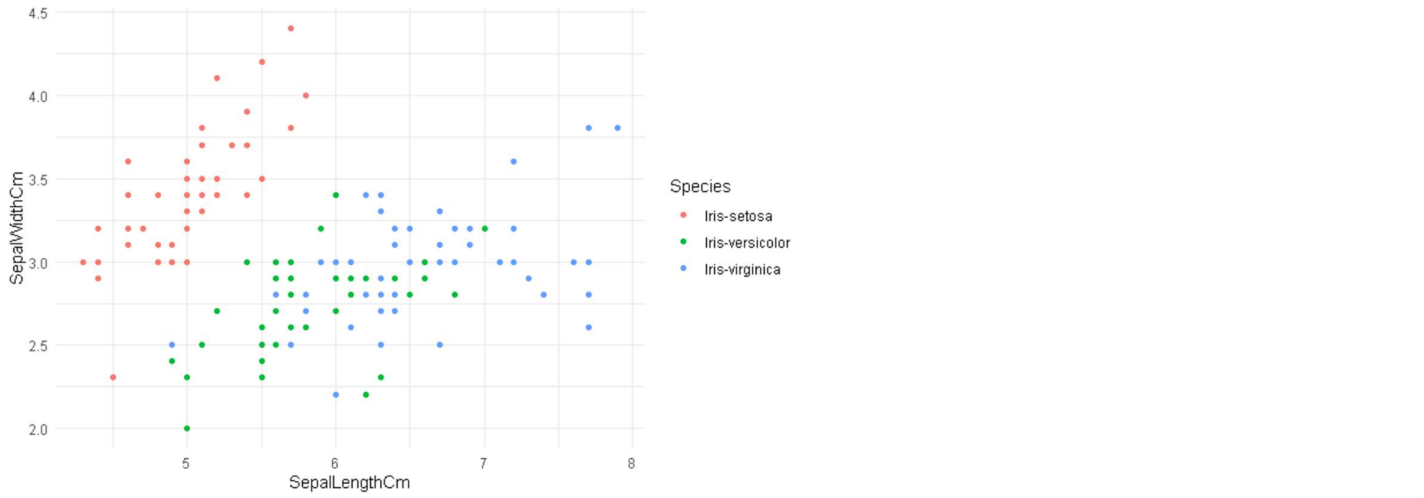
6 rows

- **use the ggplot for all the tasks below**
  Verwendung der Library "ggplot2" durch Einbindung von library('ggplot2')

````
42
43 ## use the ggplot for all the tasks below
44 - Verwendung der Libarary "ggplot2" durch Einbindung von library('ggplot2')
45
46 ```{r}
47 library(ggplot2)
48 ```
49
````

- **make a scatter plot with x-axis: Sepal.Length and y-axis: Sepal.Width, and the species should be shown in different colors**

````
49
50
51 ## make a scatter plot with x-axis: Sepal.Length and y-axis: Sepal.Width, and the species should be shown in different colors
52
53 ```{r}
54 ggplot(data=Iris, aes(x=SepalLengthCm, y=SepalWidthCm, color=Species)) + geom_point() + theme_minimal()
55 ```
````



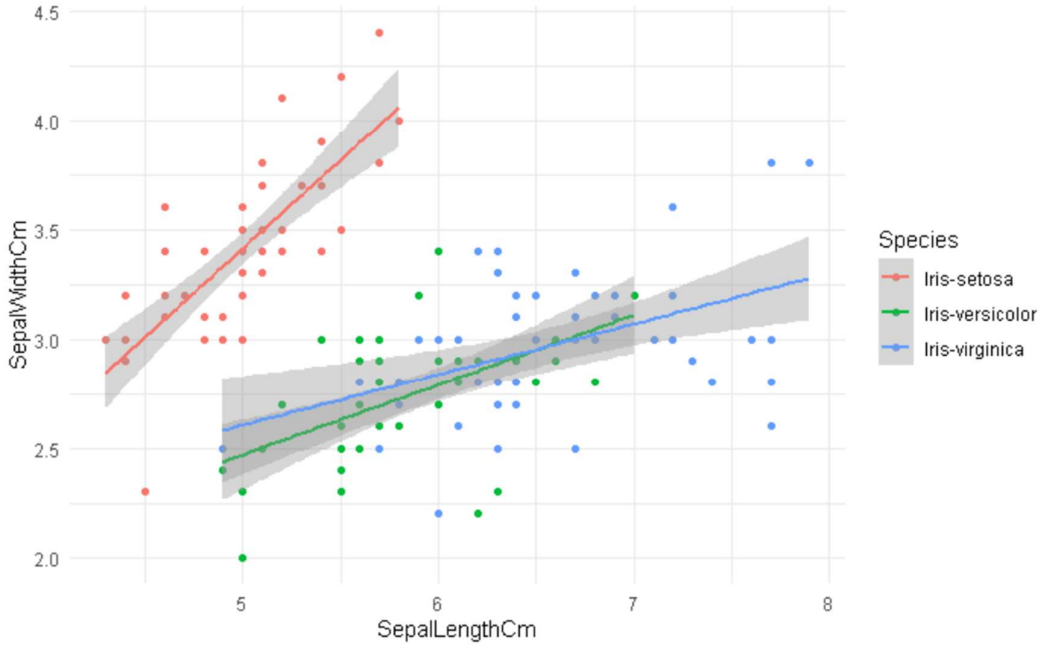
56

- **add regression line for the previous plot with the whole dataset (regardless of the species)**

```
## add regression line for the previous plot with the whole dataset (regardless of the species)

```{r}
ggplot(data=Iris, aes(x=SepalLengthCm, y=SepalWidthCm, color=Species)) + geom_point() + geom_smooth(method = lm) + theme_minimal()
```
```

```
[1m[22m`geom_smooth()` using formula = 'y ~ x'
```



- **calculate the Pearson correlation for this plot**

```
63
64
65 ## calculate the Pearson correlation for this plot
66
67 ```{r}
68 ggplot(data=Iris, aes(x=SepalLengthCm, y=SepalWidthCm, color=Species)) + geom_point() + geom_smooth(method = lm) + theme_minimal() + sm_statCorr()
69 ```
```

```
Error in sm_statCorr() : could not find function "sm_statCorr"
```

Diese Funktion kann leider nicht ausgeführt werden da es nichts ausreichend dokumentiert ist in welchen Paket es zu finden ist.

Gleiche Problematik gibt es auch für die alternative Funktion stat_cor(), welche in der offiziellen Beschreibung zu finden ist aber leider in keinen Paket wiederzufinden ist um es zu verwenden. Hier wäre es gut nochmals eine Rückmeldung zu erhalten.

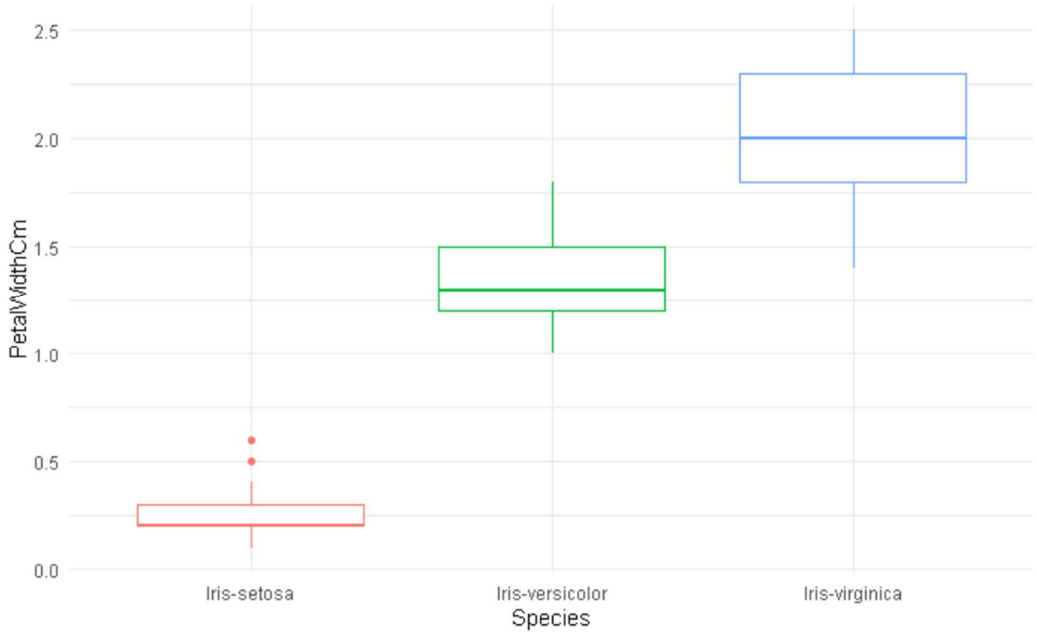
- **make the boxplot for Petal.Width with 3 species separately in the x-axis in different colors**

  Zur Ausgabe eines Boxen-Plots wird die Funktion **geom_boxplot()** verwendet.

```
## make the boxplot for Petal.Width with 3 species separately in the x-axis in different colors

```{r}
options(repr.plot.width = 5, repr.plot.height = 4)

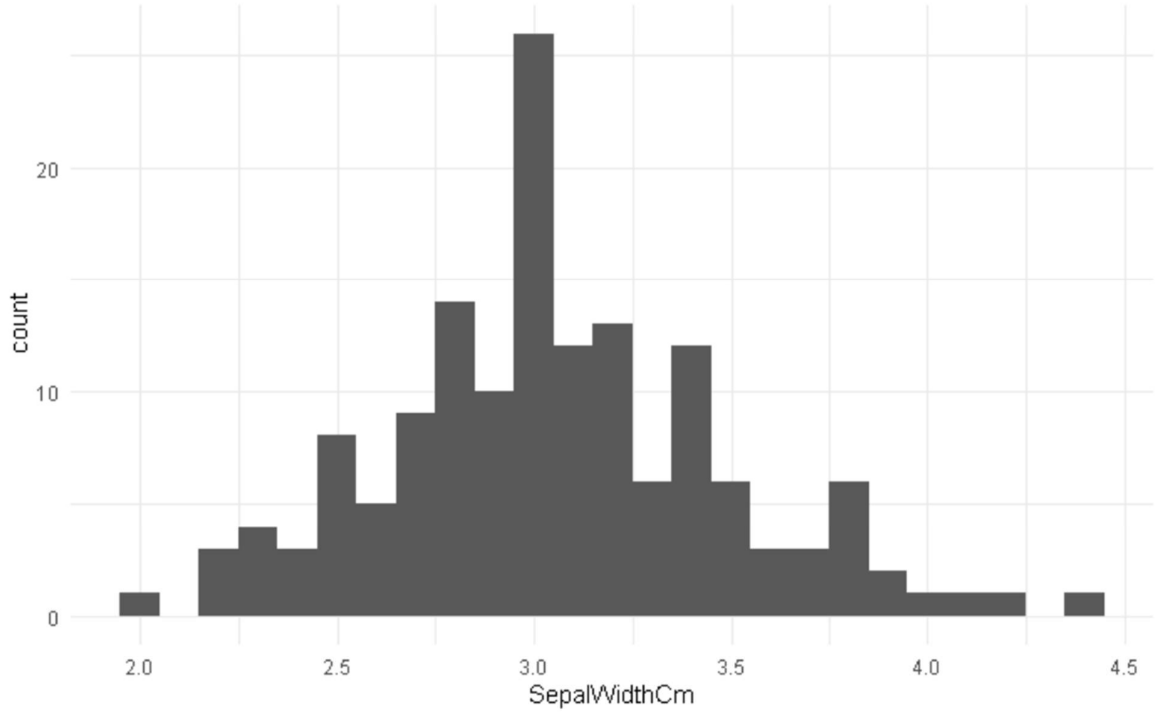
ggplot(data=Iris, aes(x=Species, y=PetalWidthCm, color=Species)) + geom_boxplot() + theme_minimal() + theme(legend.position="none")
```
```



- **make the histogram for Petal.Width with 3 species separately in x-axis in different colors**

  Zur Ausgabe eines Historgramms wird die Funktion **geom_histogram()** verwendet.

```
79
80 ## make the histogram for Petal.Width with 3 species separately in x-axis in different colors
81
82 ```{r}
83 ggplot(data=Iris,aes(x=SepalWidthCm)) + geom_histogram(binwidth=0.1) +theme_minimal()
84 ```
```




```
85
```

- **run the t-test of Petal.Width between setosa and virginica, and give the conclusion if the width is a statistically significant difference between 2 species**

````
## run the t-test of Petal.Width between setosa and virginica, and give the conclusion if the width is a statistically significant difference between 2 species

```{r}
# Subsets aufbauen, mit Filterung nach den Species-Bezeichnungen
Setosa <- subset(Iris, Species == "Iris-setosa")
Virginica <- subset(Iris, Species == "Iris-virginica")

# test ausführen
t.test(Setosa$PetalWidthCm, Virginica$PetalWidthCm)
```
````

```
        Welch Two Sample t-test

data:  Setosa$PetalWidthCm and Virginica$PetalWidthCm
t = -42.738, df = 63.594, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.865307 -1.698693
sample estimates:
mean of x mean of y
    0.244     2.026
```

Stefan Schultz, Matrikelnr. 190279 stefan.schultz@stud.th-luebeck.de TH Lübeck

Aufgabe 4: knit to HTML

Ausgabe der HTML-Datei ist auch unter dem Repo auf GitHub zu finden:

https://github.com/stefanschultz/THL_DataScience

https://github.com/stefanschultz/THL_DataScience.git