# Prettify RMD

Thomas Gaertner

February 7, 2023

## Contents



# 1 R Markdown General Introduction

- R Markdown cheat sheet
- Guide for PDF
- Guide for HTML

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

# 2 Example Analysis

In this example, we investigate a linear regression model to predict the `Sepal.Length` in the `iris` data set.

## 2.1 Data preprocessing

First, we have to load the data. Furthermore, we are transforming the variables into our target format.

```
# Loading the data
data(iris)
dat <- iris
```

As the data preprocessing might not be not so important for the interpretation, we can exclude them from the report with `include=FALSE`.

## 2.2 Some insights from the data set

First, we investigate, how many species we have. Below, you can see a table with species and their counts.

```
##
##     setosa versicolor  virginica
##         50         50         50
```

But of course, we can create a table 1 with e.g. the `qwraps2` package:

| | setosa (N = 50) | versicolor (N = 50) | virginica (N = 50) |
|---|---|---|---|
| **Sepal Length** | | | |
| Min | 4.3 | 4.9 | 4.9 |
| Mean (SD) | 5.01 ± 0.35 | 5.94 ± 0.52 | 6.59 ± 0.64 |
| Max | 5.8 | 7 | 7.9 |
| **Sepal Width** | | | |
| Min | 2.3 | 2 | 2.2 |
| Mean (SD) | 3.43 ± 0.38 | 2.77 ± 0.31 | 2.97 ± 0.32 |
| Max | 4.4 | 3.4 | 3.8 |

If you want to print a data frame, you can use `knitr::kable`, which will reformat the table nicely in the report.

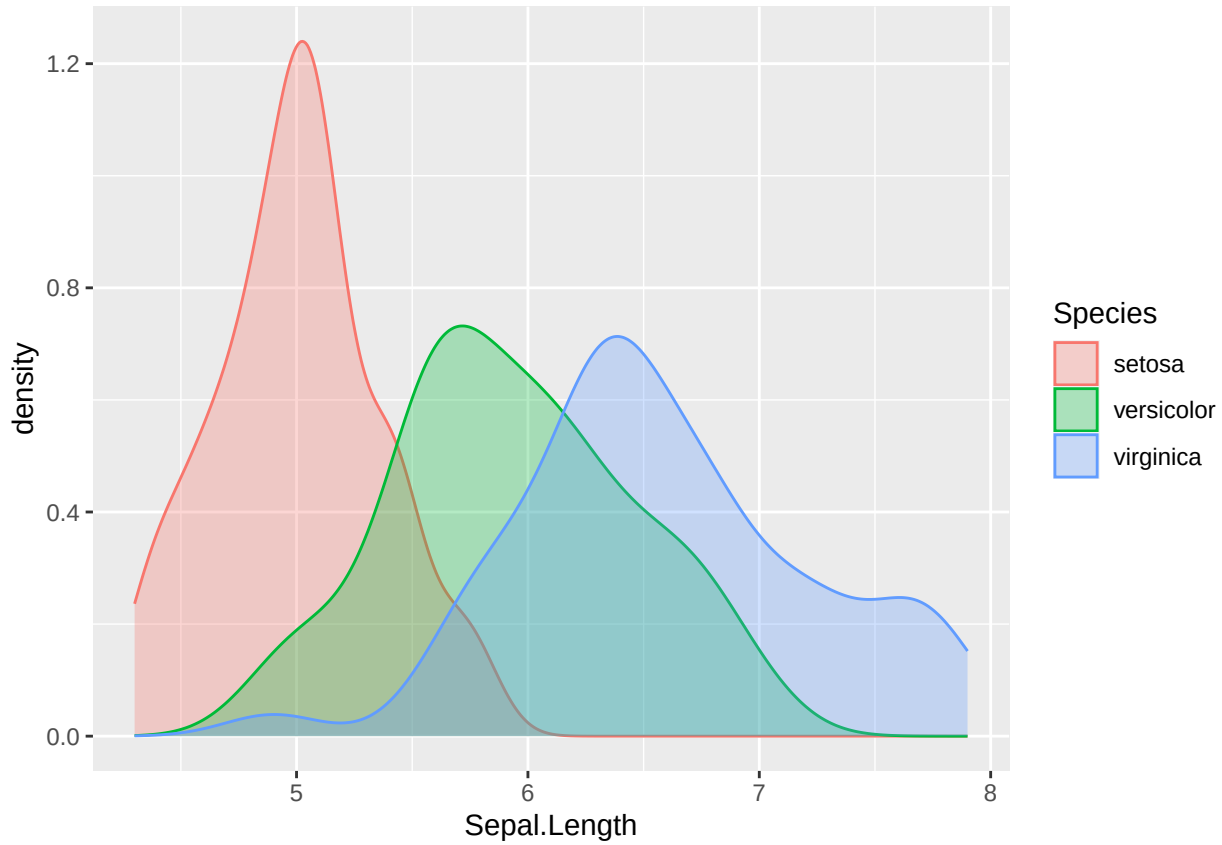
Table 2: Table with Mean Sepal Length among species

| | Mean (Sepal.Length) |
|---|---|
| setosa | 5.006 |
| versicolor | 5.936 |
| virginica | 6.588 |

Great, it seems that the sample mean is different between the groups. [1]

In the next step, we are looking into the density plots. We will create them with `ggplot2`.

[1] This is not a formal test. For comparing the mean, a t-test should be applied.




Cool, we have included a nice graph.

## 2.3 Linear Model

In the next step, we will fit a linear model, which is estimating the `Sepal.Length`. As variables, we include the `species`, the `petal` values and the `Sepal.Width`. We will also print the summary. We assume for the analysis a significance level of $\alpha = 0.05$

```
fit <- lm(Sepal.Length ~ Species + Sepal.Width, data = dat)
summary(fit)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species + Sepal.Width, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30711 -0.25713 -0.05325  0.19542  1.41253
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.2514     0.3698   6.089 9.57e-09 ***
## Speciesversicolor   1.4587     0.1121  13.012  < 2e-16 ***
## Speciesvirginica    1.9468     0.1000  19.465  < 2e-16 ***
## Sepal.Width         0.8036     0.1063   7.557 4.19e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.438 on 146 degrees of freedom
## Multiple R-squared:  0.7259, Adjusted R-squared:  0.7203
## F-statistic: 128.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

We can also print out a single output from our summary, if we want. We can call for instance in the summary some values. For that, we are using `print` and `paste` (for string concatenation).

```
## [1] "R-Squared:  0.726"
```

And again, we are including a nice table of the coefficients with `knitr::kable`.

Table 3: Coefficients table from `fit`.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.2513932 | 0.3697543 | 6.088890 | 0 |
| Speciesversicolor | 1.4587431 | 0.1121079 | 13.011954 | 0 |
| Speciesvirginica | 1.9468166 | 0.1000150 | 19.465255 | 0 |
| Sepal.Width | 0.8035609 | 0.1063390 | 7.556598 | 0 |

**Findings (some example interpretations)**

1) The model has a p-value of $<$ 2.2e-16, which is less than our significance level of $\alpha = 0.05$.
2) The R-squared values is 0.726, so ~73% of the variance of sepal length can be explained by our model.
3) All coefficients has a p-value $< \alpha$ and are significant.
4) The sepal length of species "versicolor" are expected to be 1.45 larger than the species setosa.
5) With a sepal width increase of 1, we can expect an increase of 0.8 for the sepal length.