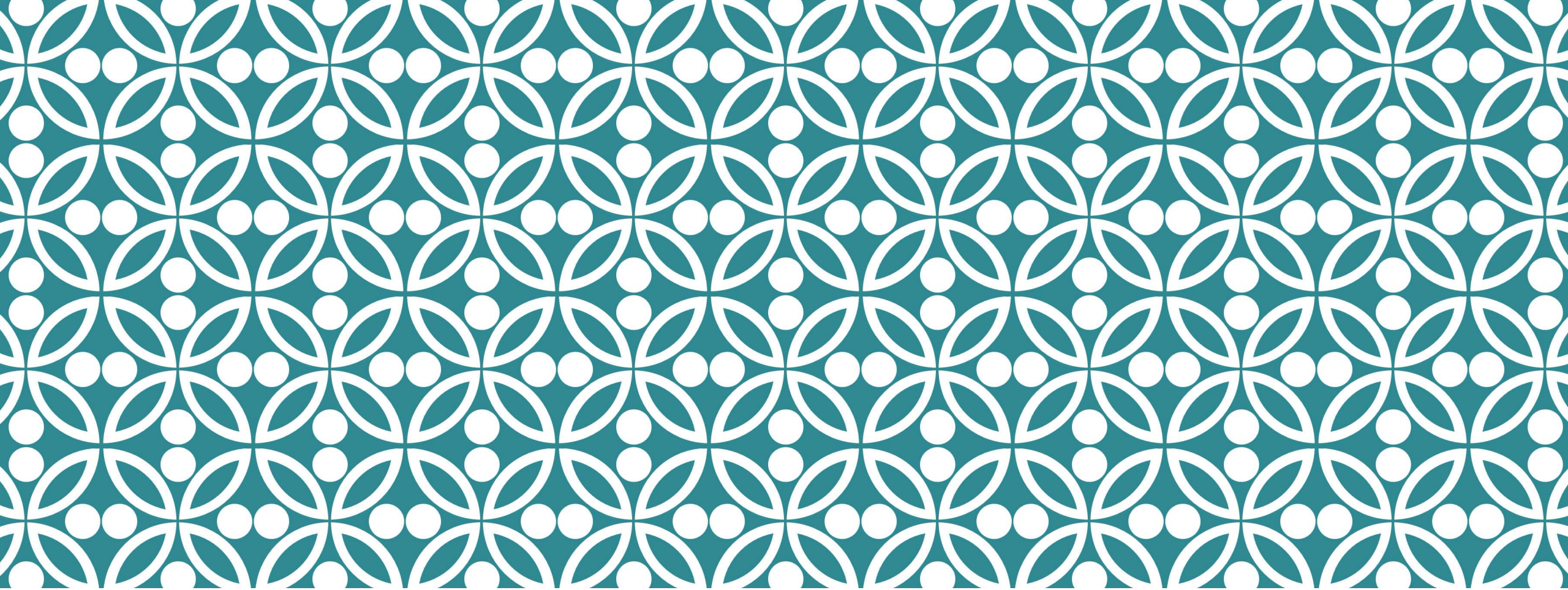


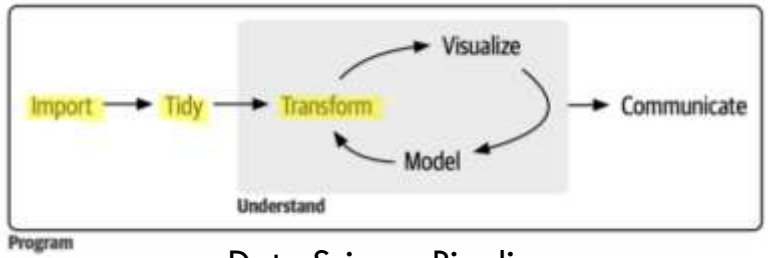
SESSION 3: STATISTICS

DR. SOFIA GIL-CLAVEL

- Recap
- Fundamentals of modeling in R: Example applied to linear models
- Other models?

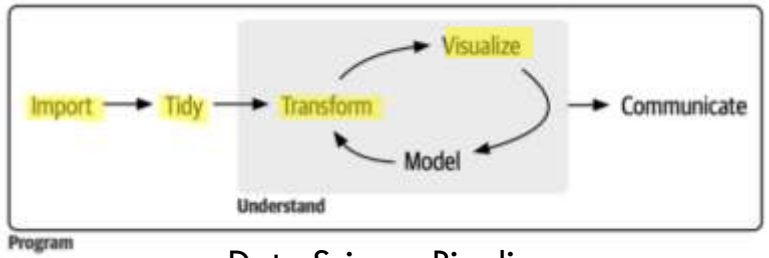
1. SESSION 1-2: RECAP

SESSION 1: DATA MANAGEMENT



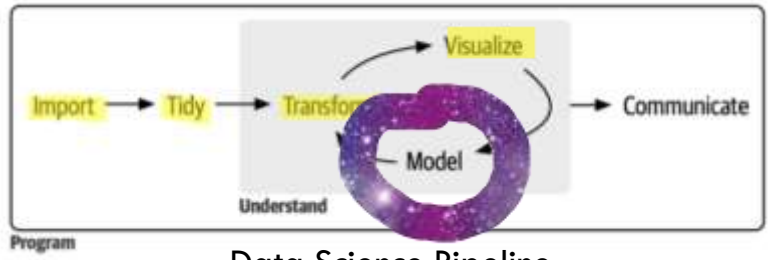
Data Science Pipeline

SESSION 2: DATA VISUALIZATION



Data Science Pipeline

SESSION 3: STATISTICS



Data Science Pipeline

stats-package {stats} R Documentation

The R Stats Package

Description

R statistical functions

Details

This package contains functions for statistical calculations and random number generation.

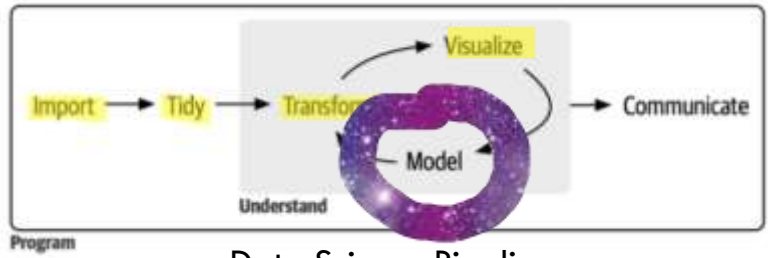
For a complete list of functions, use `library(help = "stats")`.

Author(s)

R Core Team and contributors worldwide

Maintainer: R Core Team R-core@r-project.org

SESSION 3: STATISTICS



Data Science Pipeline

stats-package {stats} R Documentation

The R Stats Package

Description

R statistical functions

Details

This package contains functions for statistical calculations and random number generation.

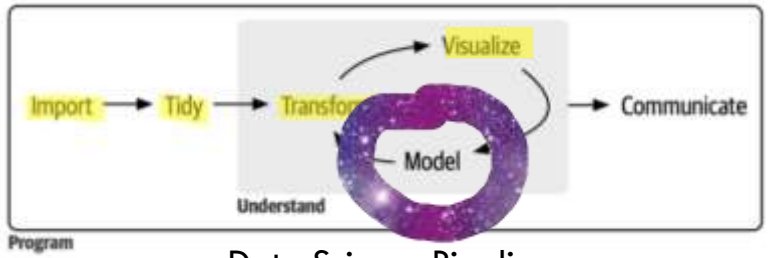
For a complete list of functions, use `library(help = "stats")`.

Author(s)

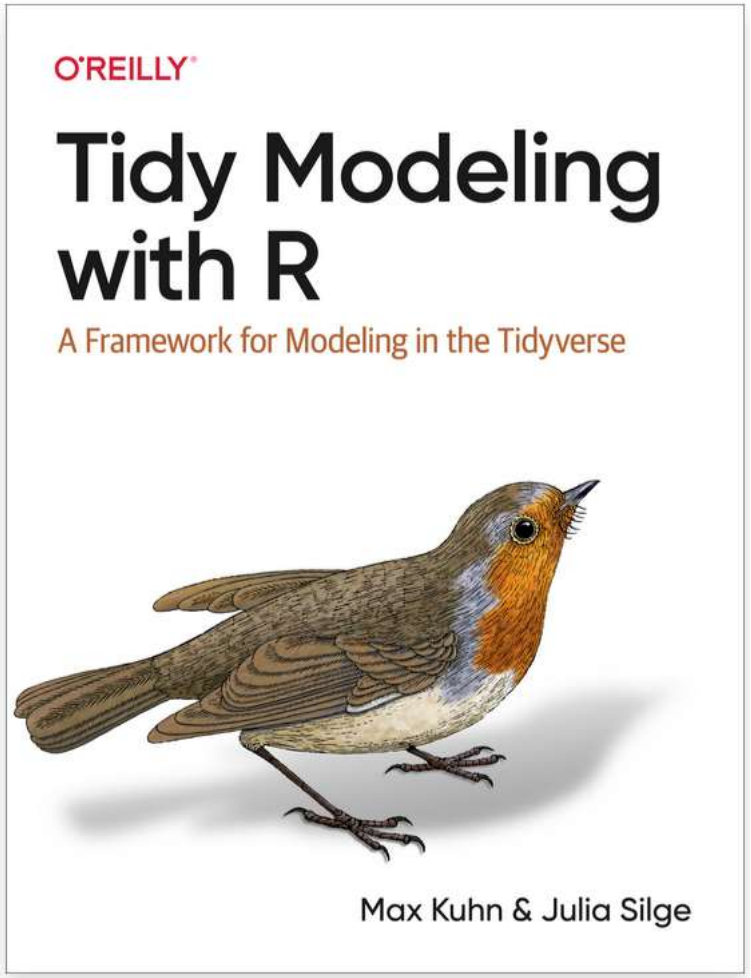
R Core Team and contributors worldwide

Maintainer: R Core Team R-core@r-project.org

SESSION 3: STATISTICS



Data Science Pipeline



Check for free here: https://www.tmwr.org/

stats-package {stats} R Documentation

The R Stats Package

Description

R statistical functions

Details

This package contains functions for statistical calculations and random number generation.

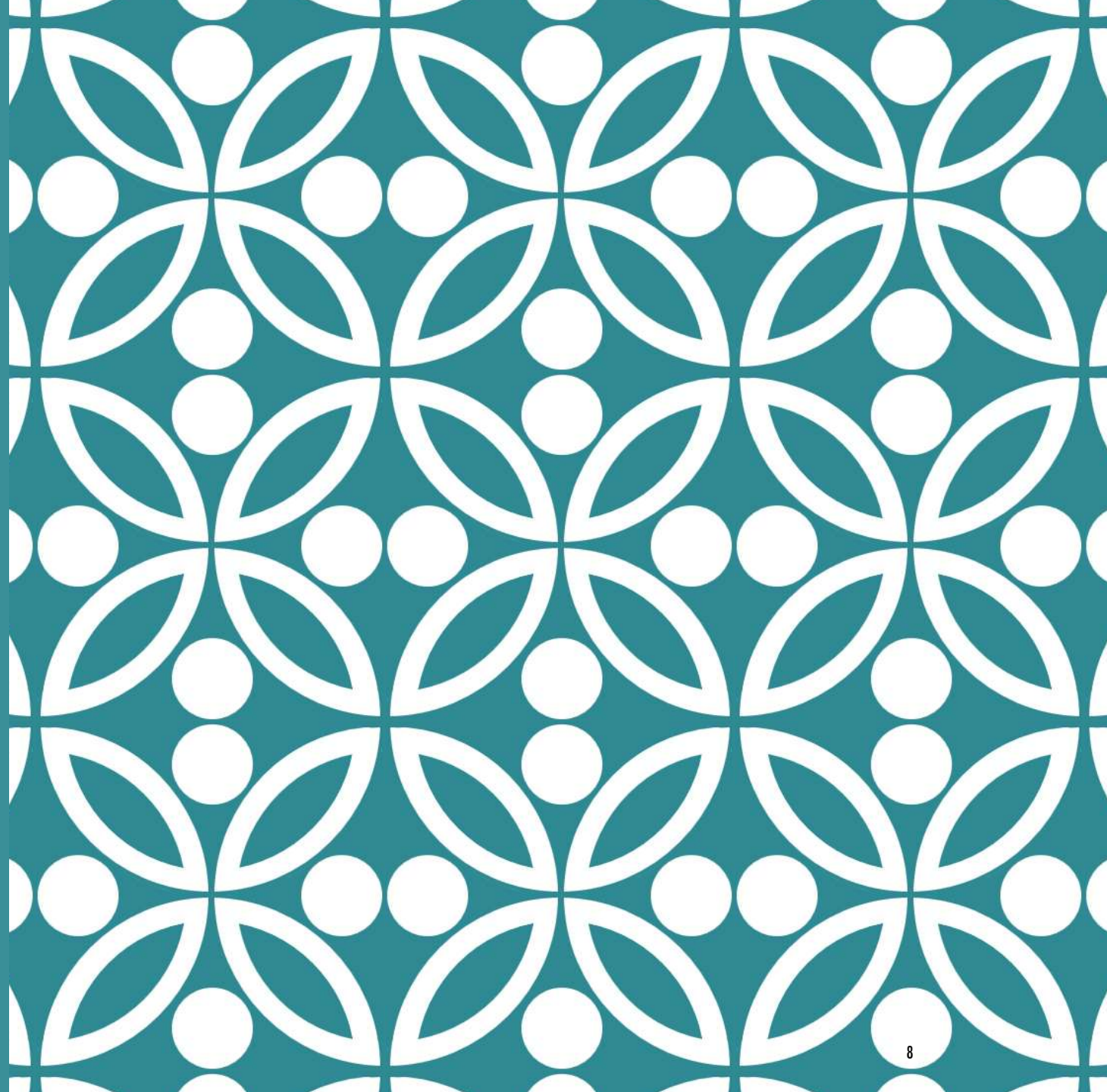
For a complete list of functions, use `library(help = "stats")`.

Author(s)

R Core Team and contributors worldwide

Maintainer: R Core Team R-core@r-project.org

3. FUNDAMENTALS OF R-MODELING



TYPES OF VARIABLES

- **Dependent**: The variable (also known as outcome) we are analyzing and that we believe its behavior "depends" on other variables.

$$Y \sim$$

TYPES OF VARIABLES

- **Dependent**: The variable (also known as outcome) we are analyzing and that we believe its behavior "depends" on other variables.

- **Independent**: The variables to which the "dependent" variables is tied to.

$$Y \sim X1 + X2 + X3$$

TYPES OF VARIABLES

- **Dependent**: The variable (also known as outcome) we are analyzing and that we believe its behavior "depends" on other variables.

- **Independent**: The variables to which the "dependent" variables is tied to.

Symbolic Language

$$Y \sim X1 + X2 + X3$$

TYPES OF VARIABLES

- **Dependent**: The variable (also known as outcome) we are analyzing and that we believe its behavior "depends" on other variables.

- **Independent**: The variables to which the "dependent" variables is tied to.

Symbolic Language

$$Y \sim X1 + X2 + X3$$

 was built to understand symbolic language!

BUT FIRST, LET'S OPEN SOME DATA

For this section, we will use the database "**ChickWeight**" from the basic R package.

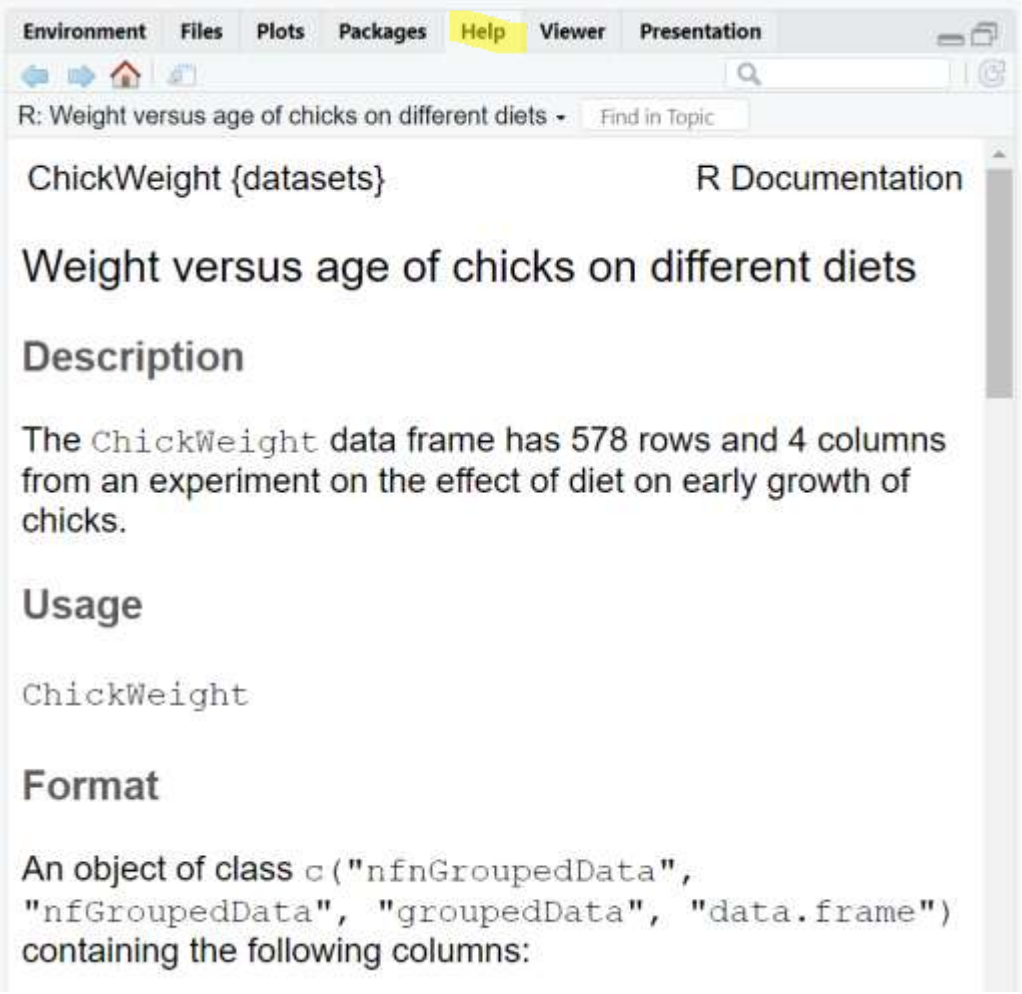
To open it, you just need to write down the name in the console:

```
ChickWeight
```

To see in an independent window, you can use:

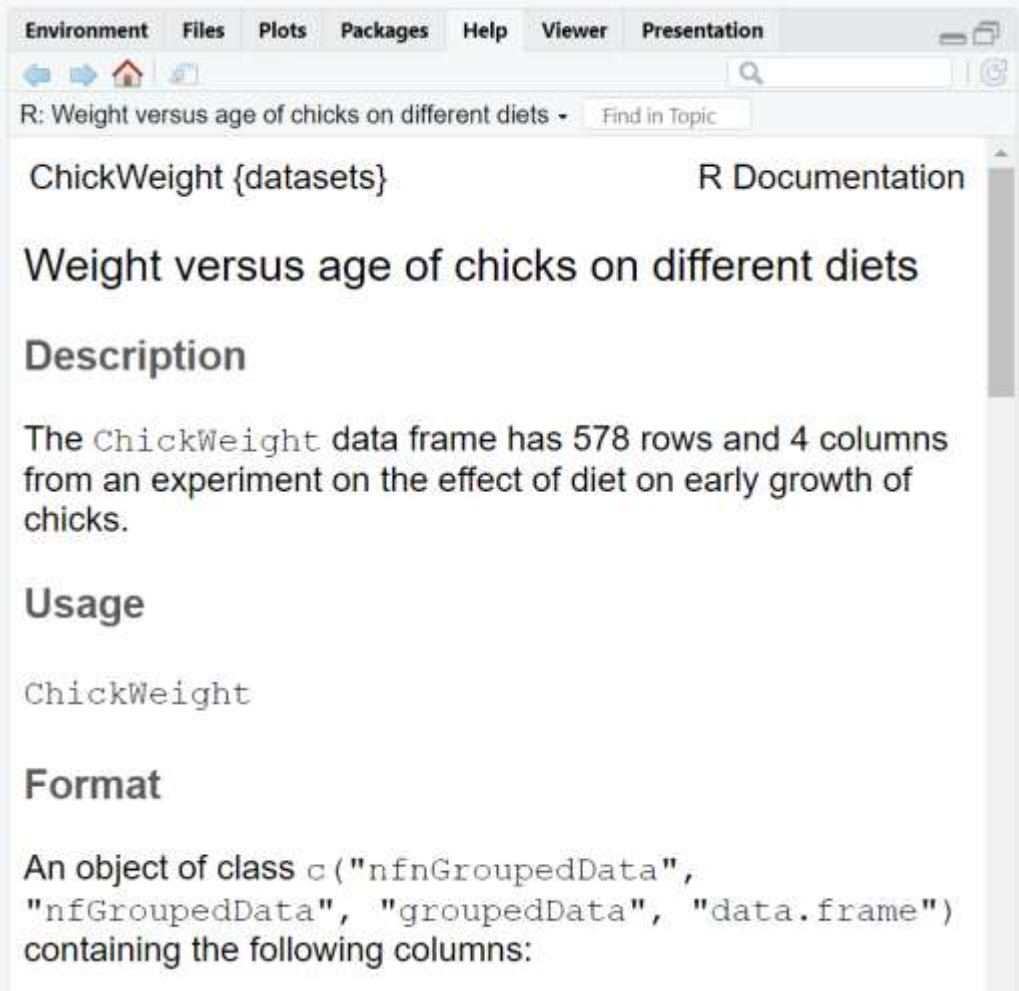
```
View(ChickWeight)
```

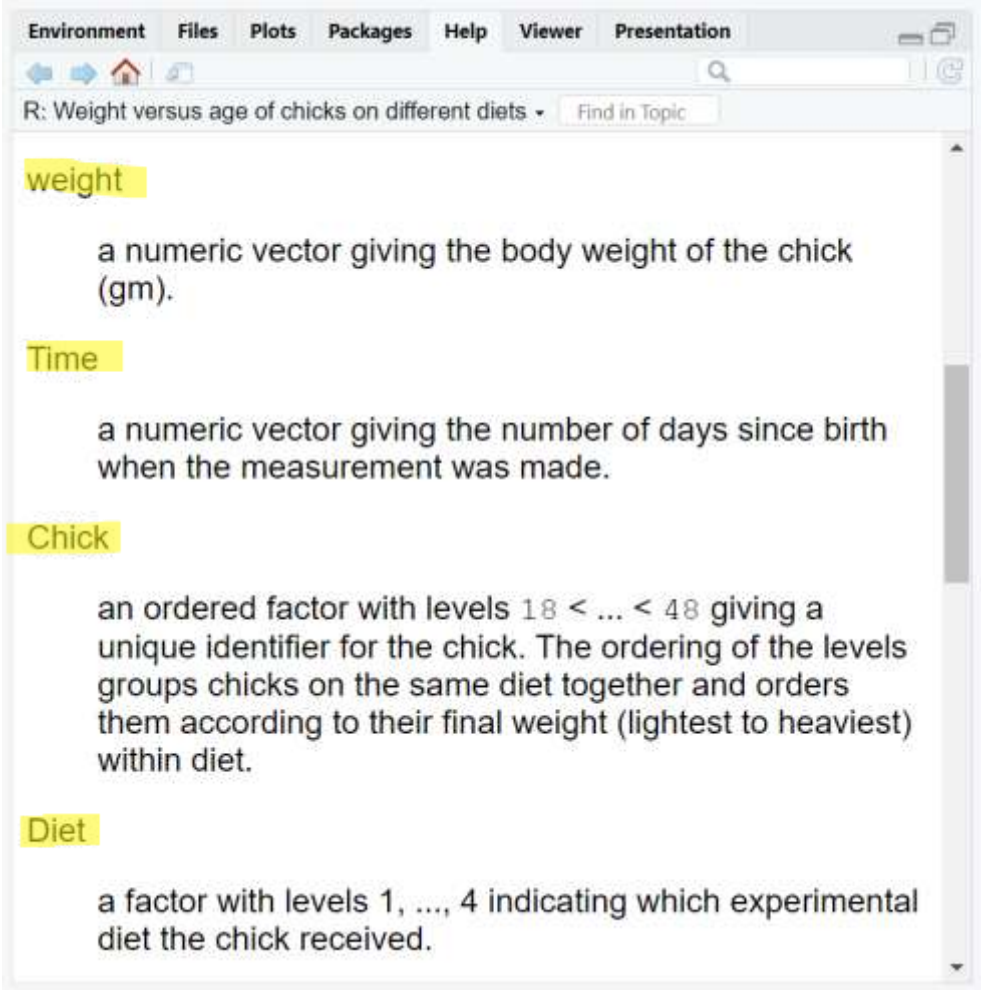
BUT FIRST, LET'S OPEN SOME DATA



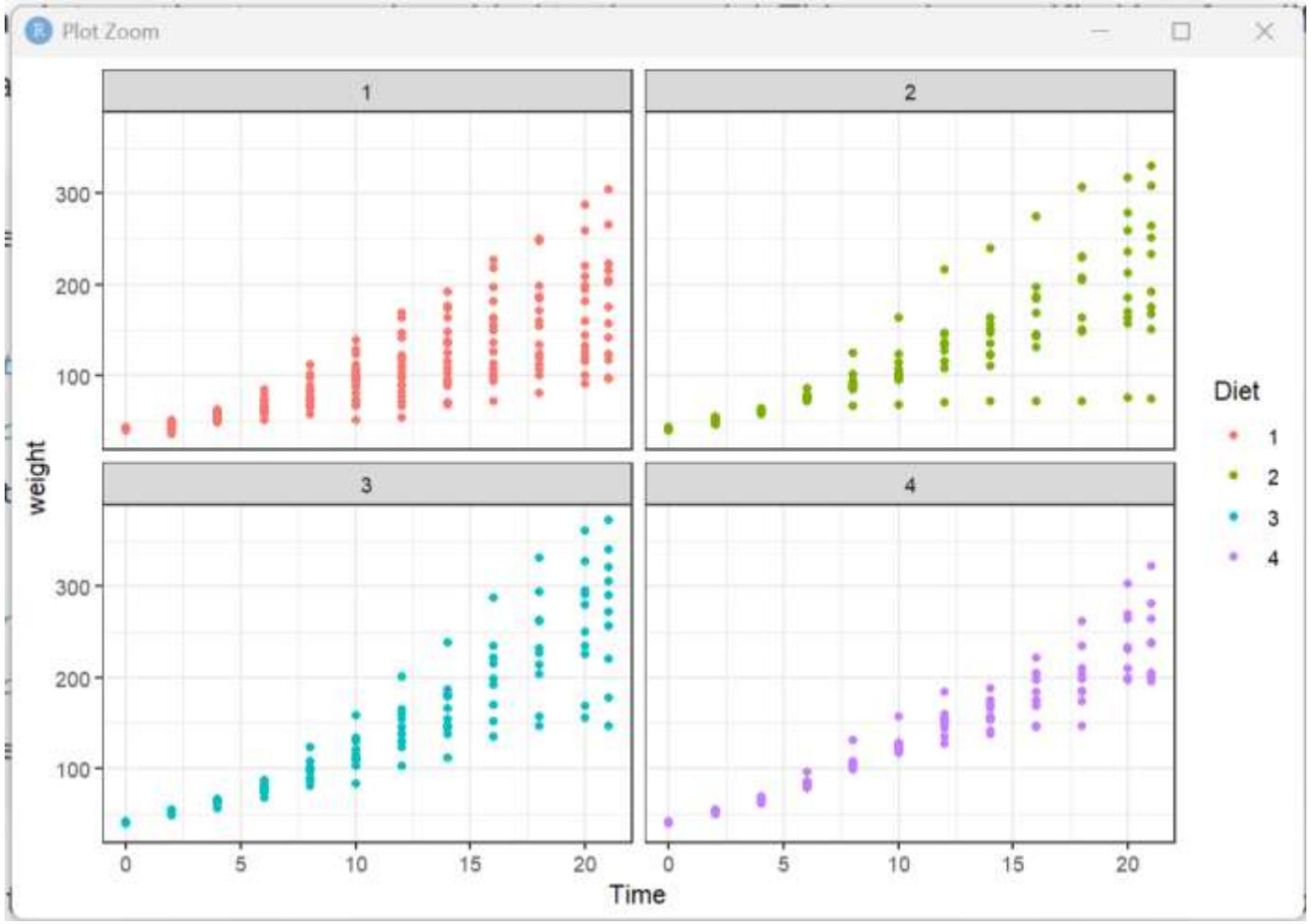
To learn what each variable represents, you can check its documentation in the "help" window.

BUT FIRST, LET'S OPEN SOME DATA

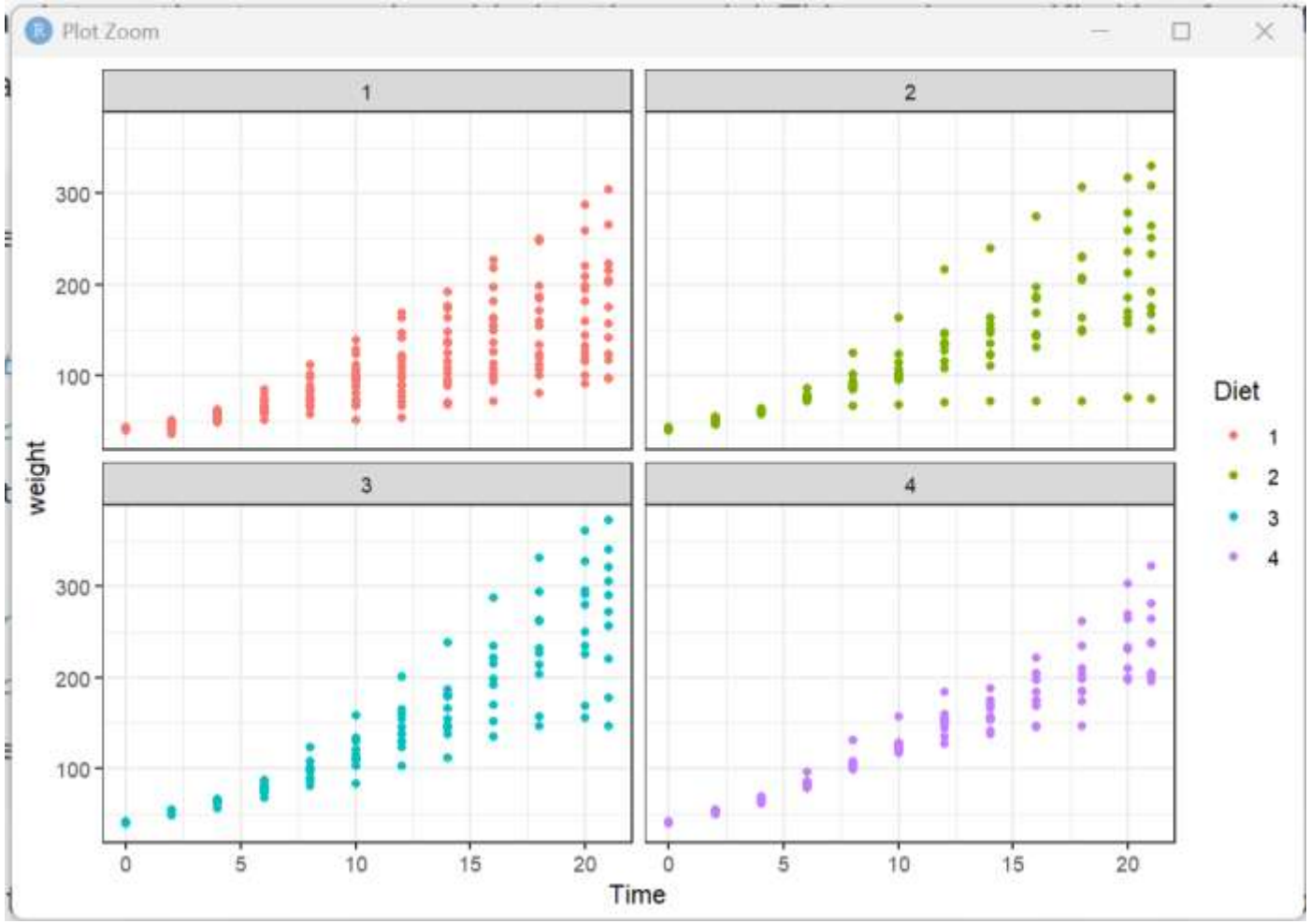




LET'S CHECK HOW THE DATA LOOKS

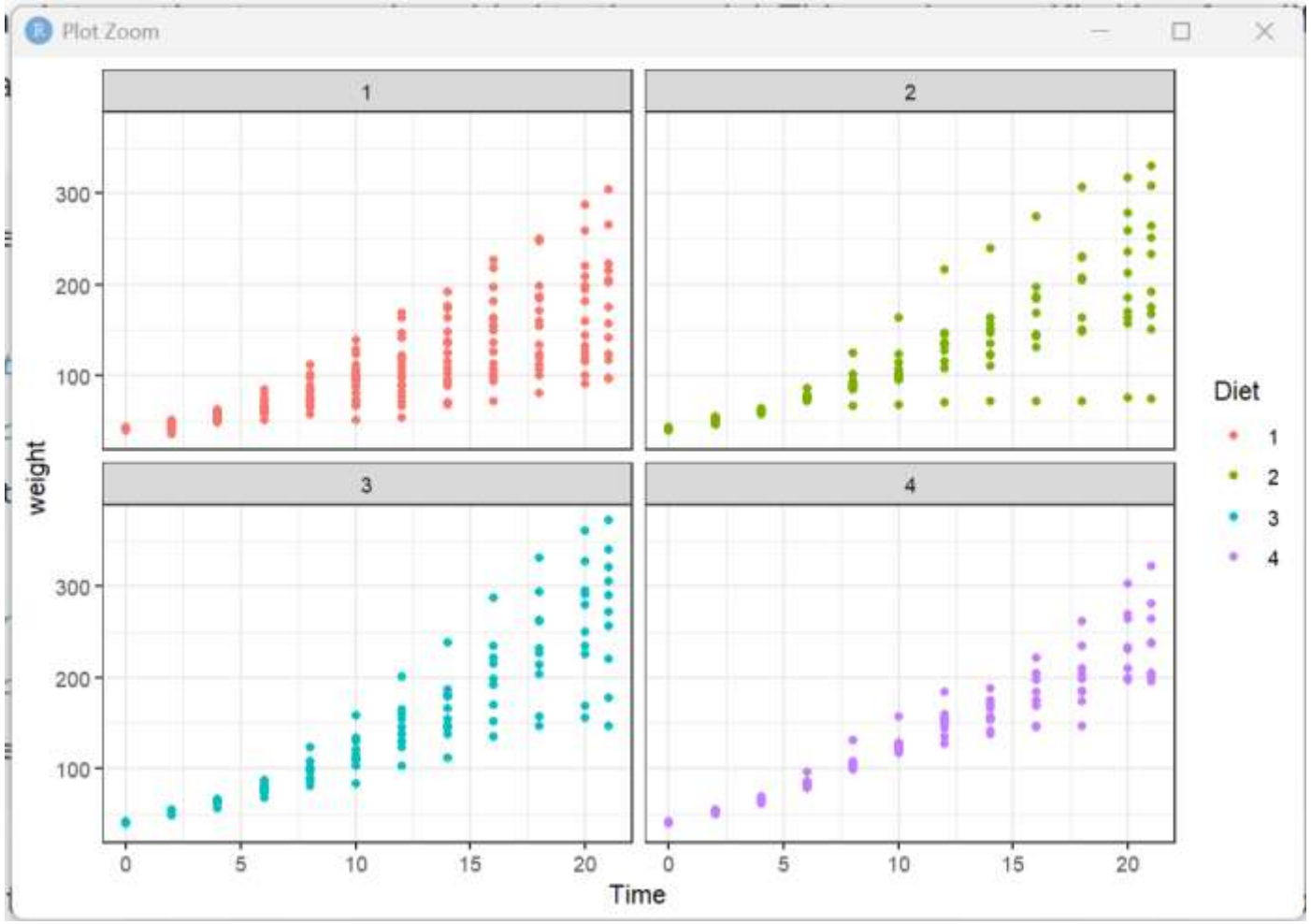


LET'S CHECK HOW THE DATA LOOKS



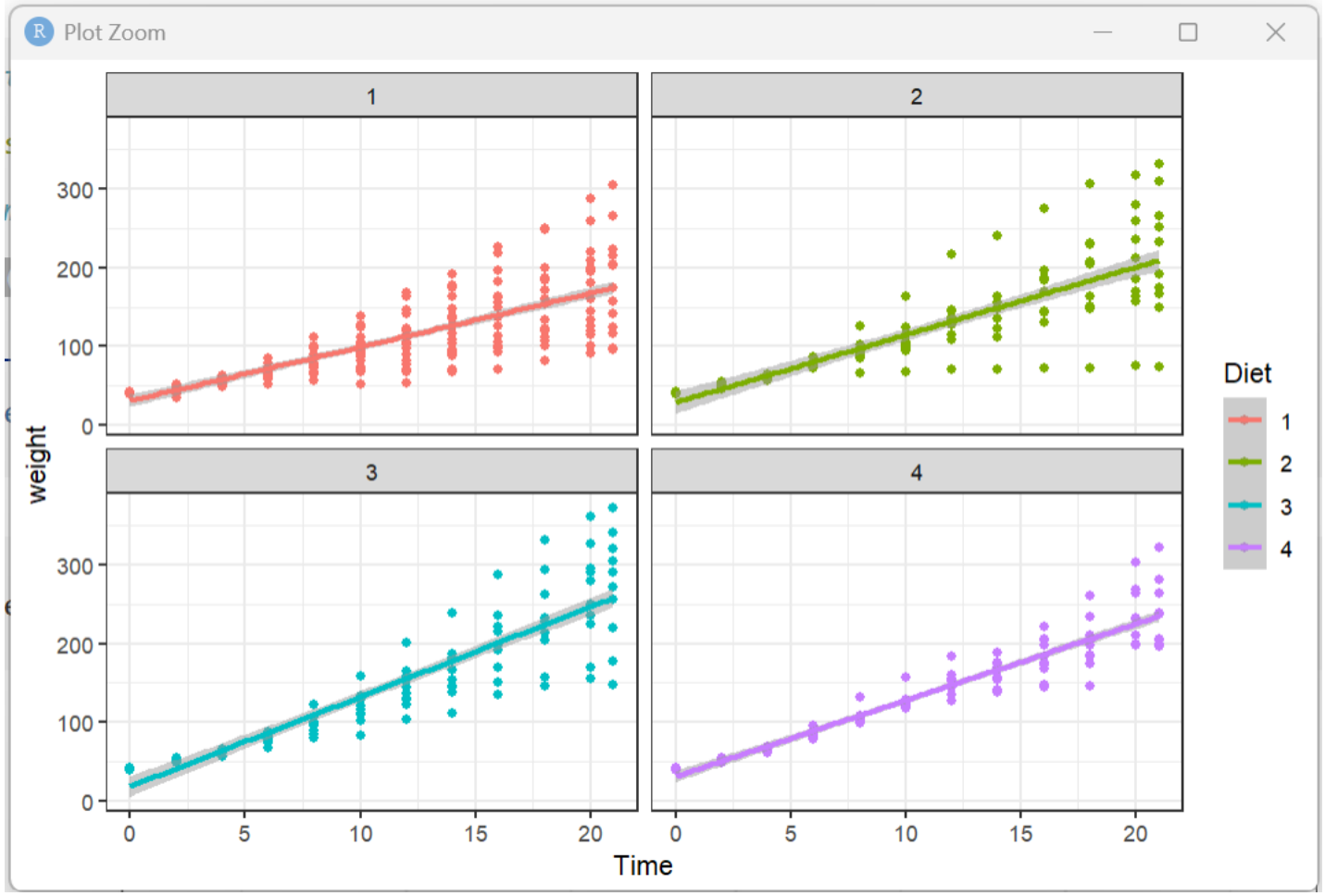
What are some interesting statistical questions?

LET'S CHECK HOW THE DATA LOOKS



What are some interesting statistical questions? That is up to you!

LET'S CHECK HOW THE DATA LOOKS

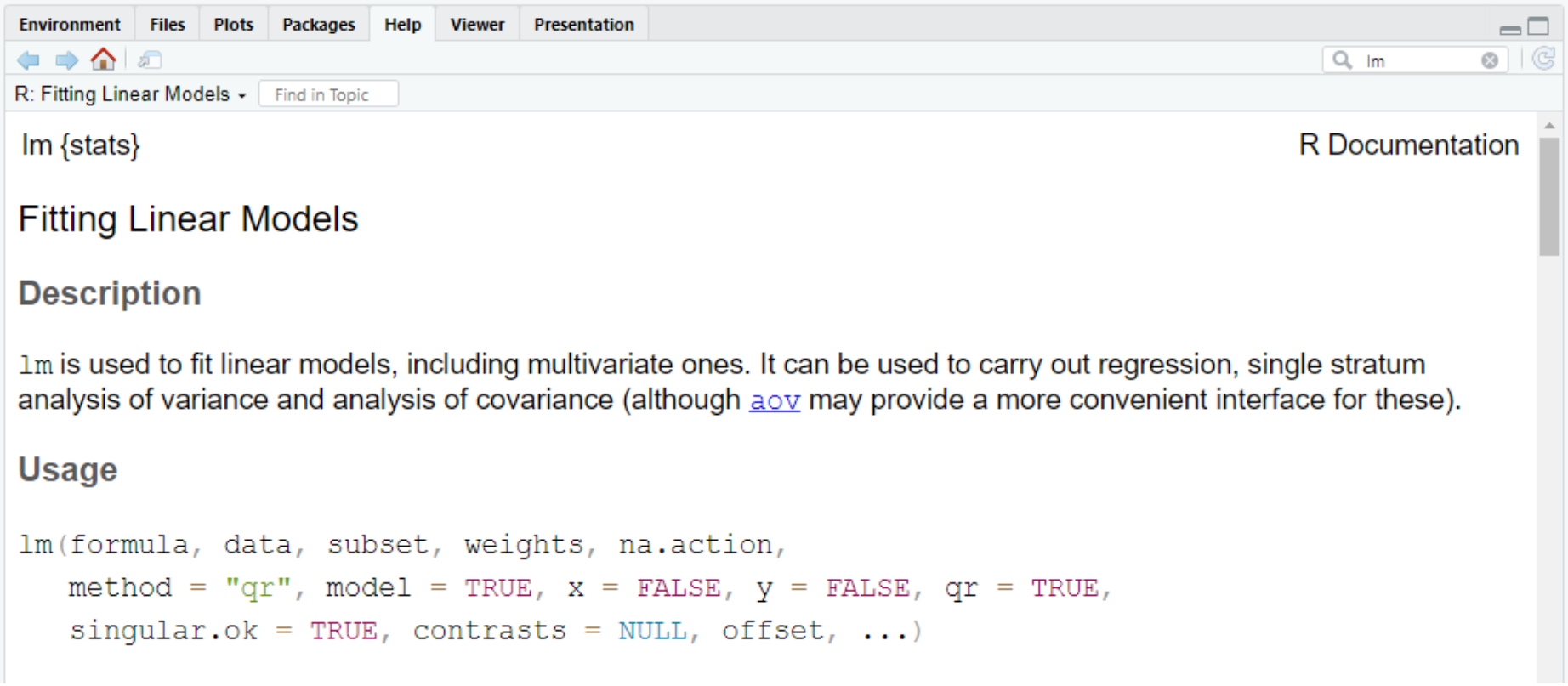


What are some interesting statistical questions? That is up to you!

Now we will learn about:

- Linear regression models

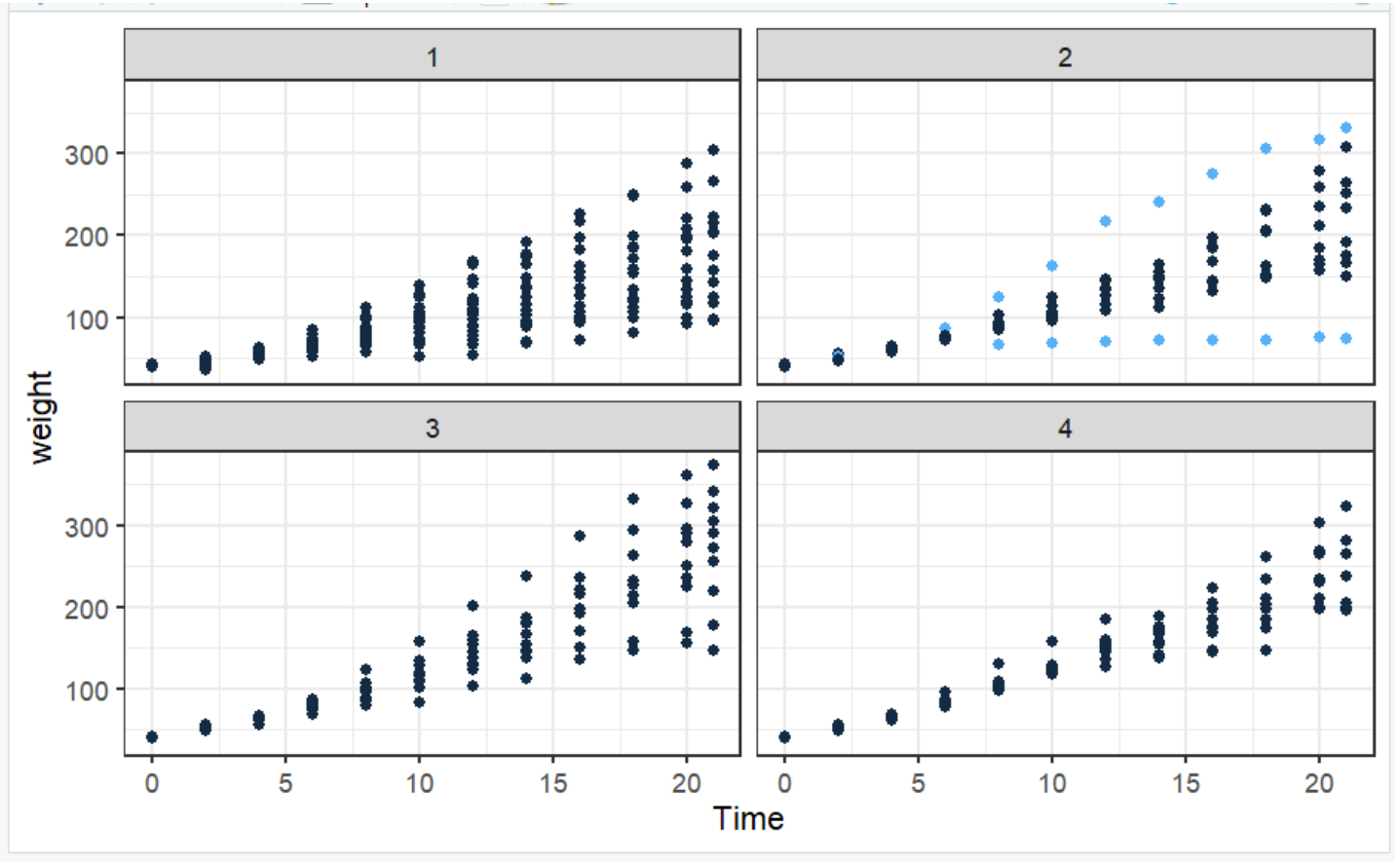
FITTING LINEAR MODELS



HANDLING OUTLIERS

An outlier can be higher or lower than expected or **displaced more to the right or left than expected.**

Outliers can affect regression lines, **making the regression lines less accurate** in predicting other data.



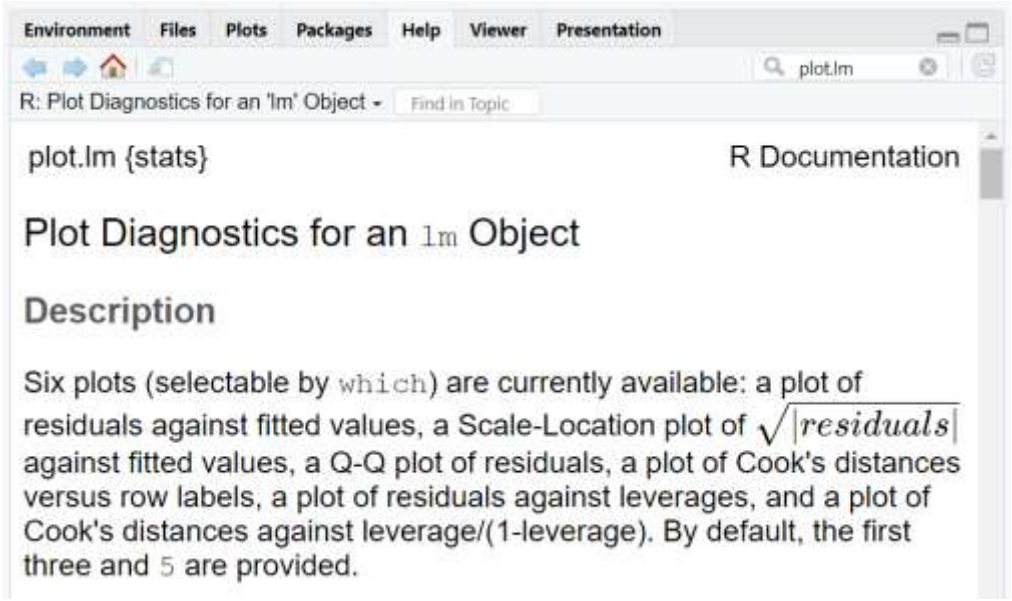
CHECKING THE ASSUMPTIONS

There are five key assumptions of linear regression:

1. 📈 **Linearity**
2. 🔵 **independence**
3. 📊 **homoscedasticity**
4. 🔔 **normality**
5. 🚫 **no multicollinearity**

Ensuring these assumptions are met is critical to creating an accurate and reliable model for predicting and drawing insights from data.

CHECKING THE ASSUMPTIONS



CHECKING THE ASSUMPTIONS

Environment | Files | Plots | Packages | Help | Viewer | Presentation

plot.lm

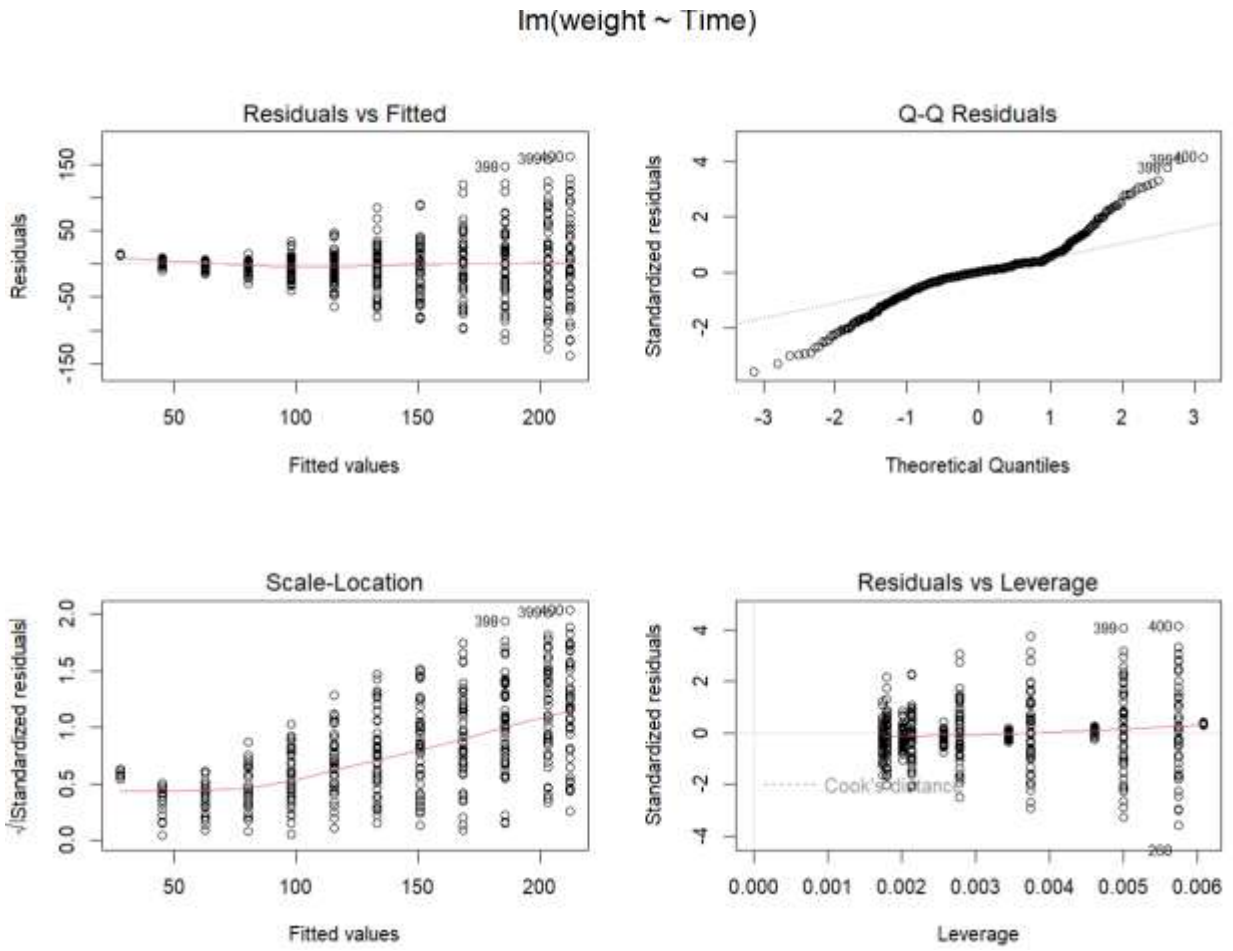
R: Plot Diagnostics for an 'lm' Object — Find in Topic

plot.lm {stats} R Documentation

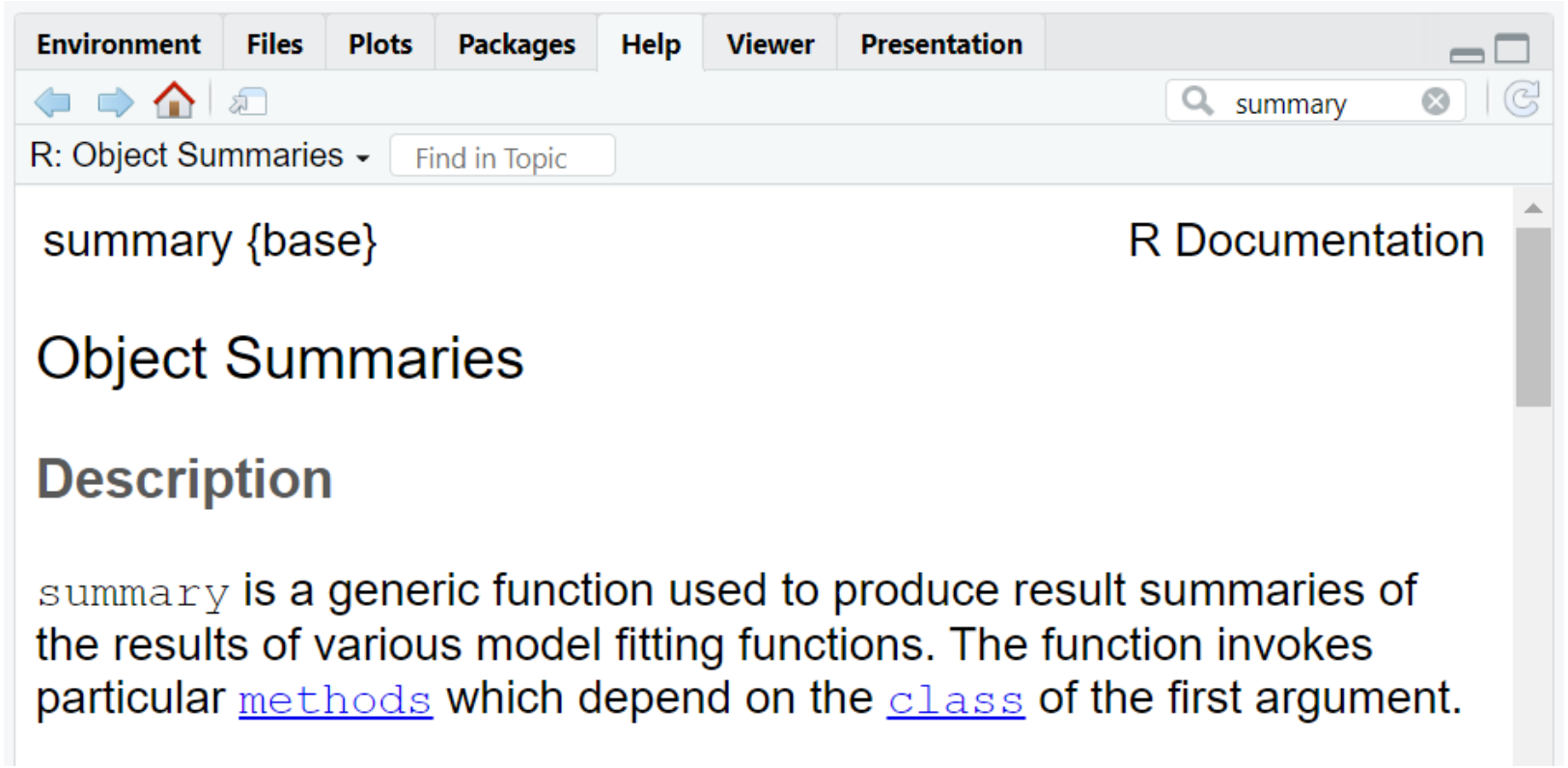
Plot Diagnostics for an `lm` Object

Description

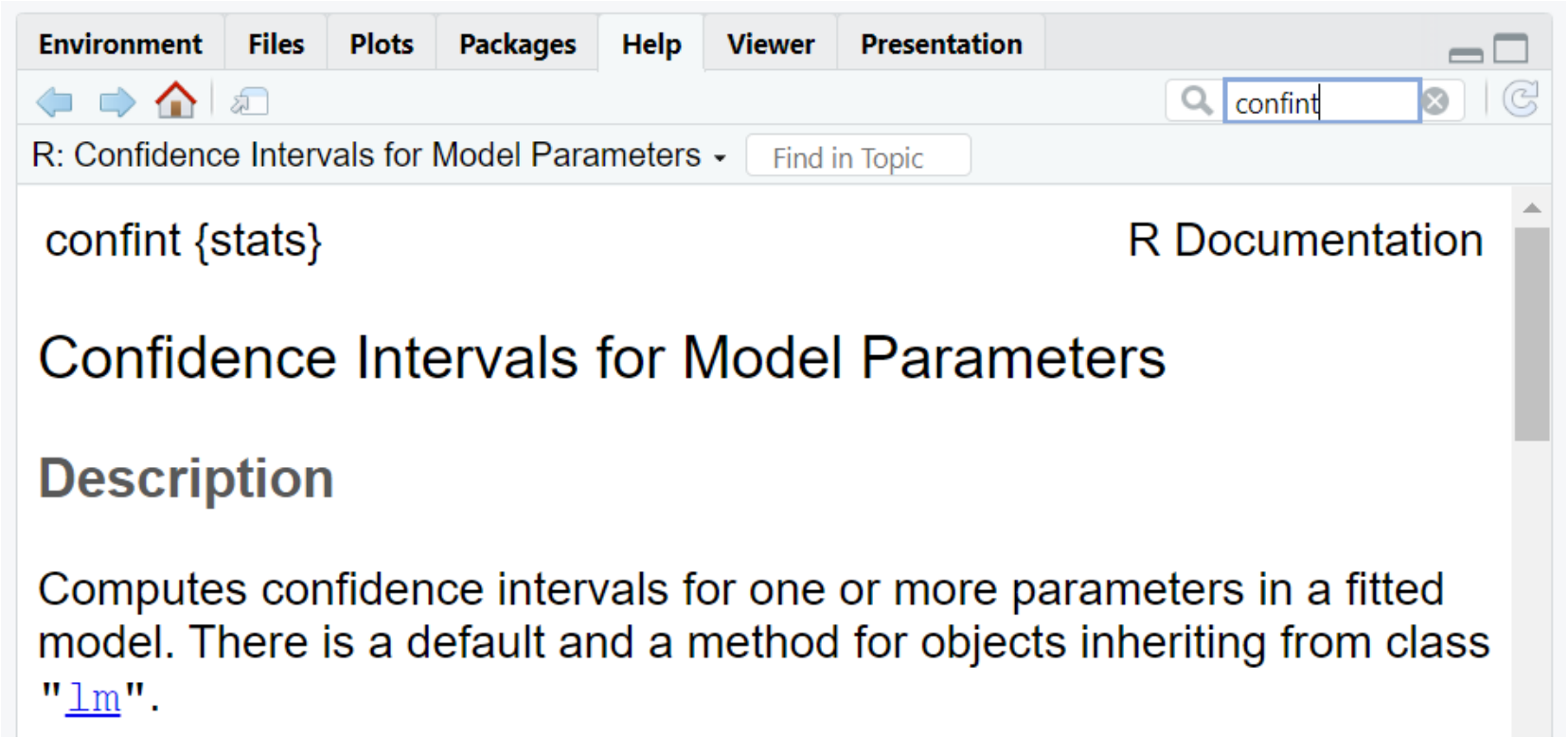
Six plots (selectable by `which`) are currently available: a plot of residuals against fitted values, a Scale-Location plot of $\sqrt{|residuals|}$ against fitted values, a Q-Q plot of residuals, a plot of Cook's distances versus row labels, a plot of residuals against leverages, and a plot of Cook's distances against leverage/(1-leverage). By default, the first three and `5` are provided.



CHECK THE COEFFICIENTS AND OTHER STATS



CONFIDENCE INTERVALS



INTERACTION TERMS

Statistical **interaction** means **the effect of one independent variable**(s) **on the dependent variable depends on the value of another independent variable**(s).

INTERACTION TERMS

Statistical **interaction** means **the effect of one independent variable**(s) **on the dependent variable depends on the value of another independent variable**(s).

$$Y \sim X1 \times X2$$

INTERACTION TERMS

Statistical **interaction** means **the effect of one independent variable**(s) **on the dependent variable depends on the value of another independent variable**(s).

$$Y \sim X1 \times X2$$

So, in our data, we could think that **the effect that time has on the chickens' weight depends on the diet**:

$$weight \sim Time \times Diet$$

INTERACTION TERMS

Statistical **interaction** means **the effect of one independent variable**(s) **on the dependent variable depends on the value of another independent variable**(s).

$$Y \sim X1 \times X2$$

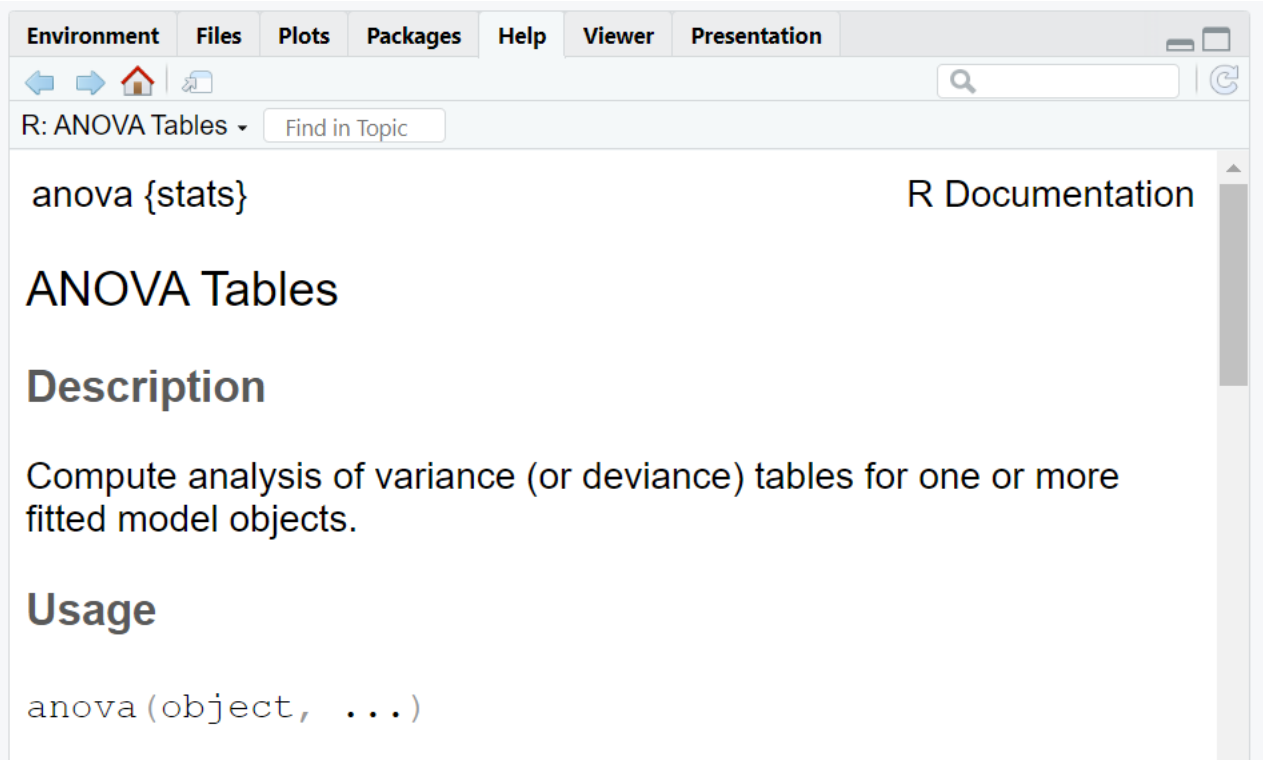
So, in our data, we could think that **the effect that time has on the chickens' weight depends on the diet:**

$$weight \sim Time \times Diet$$

Remember, this is **Symbolic Language**

ANOVA: ANALYSIS OF VARIANCE

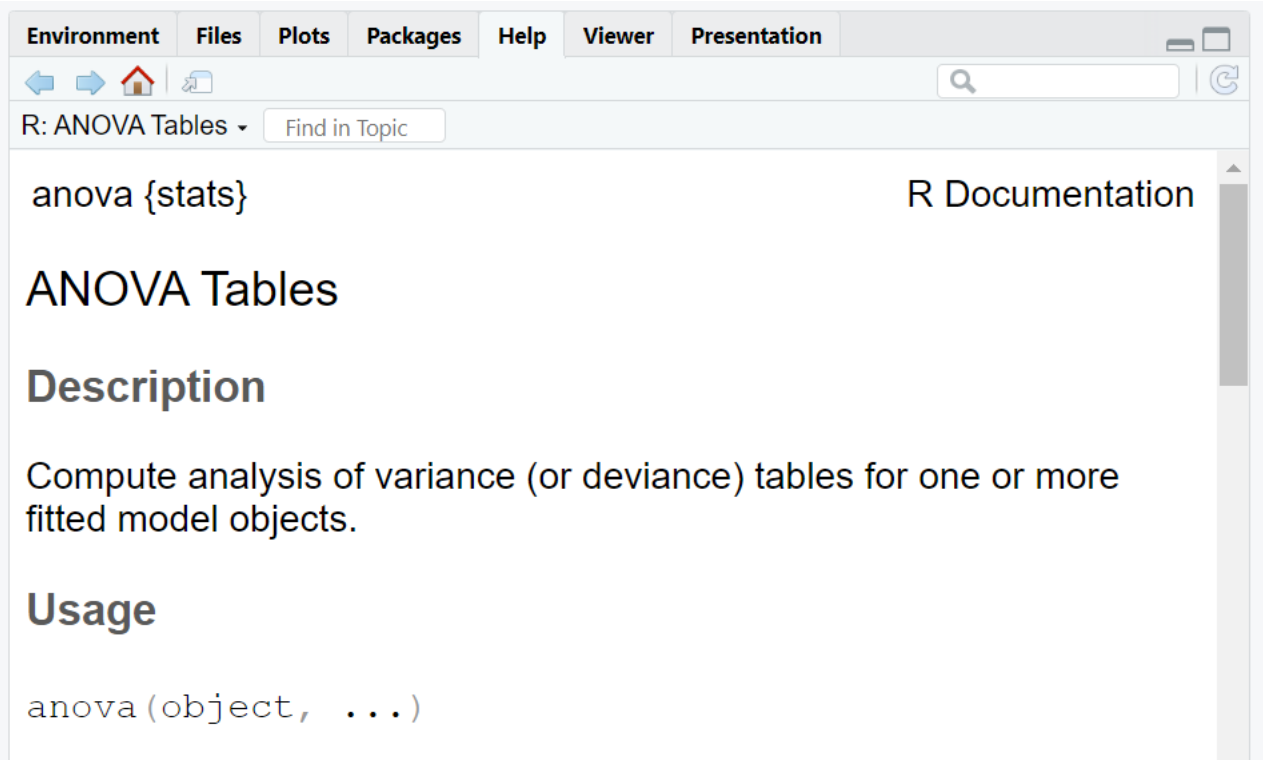
The ANOVA test is very useful when we want to compare two models to see which one fits the data better.



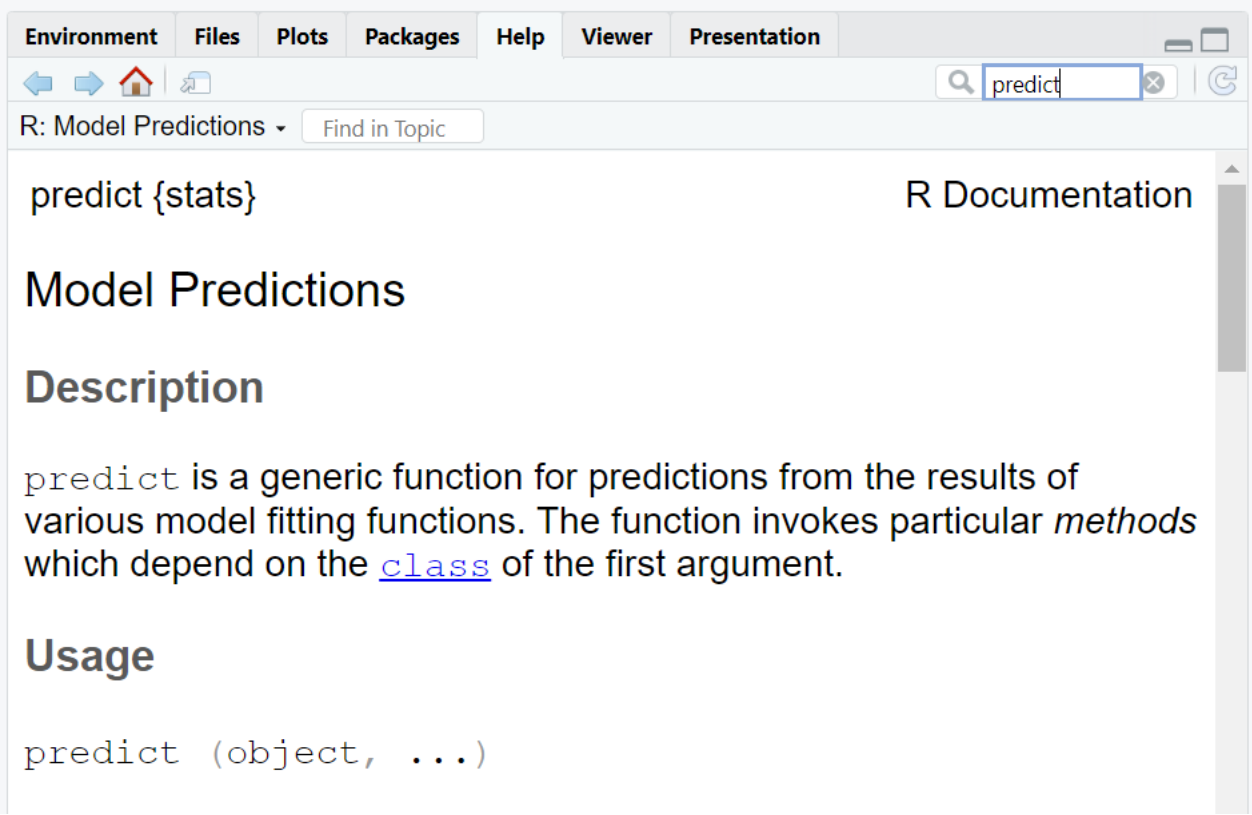
ANOVA: ANALYSIS OF VARIANCE

The ANOVA test is very useful when we want to compare two models to see which one fits the data better.

So, which model is better, with or without interaction term?



PREDICTING


Environment
Files
Plots
Packages
Help
Viewer
Presentation
predict
R: Model Predictions
Find in Topic
predict {stats}
R Documentation
Model Predictions
Description
predict is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument.
Usage
predict (object, ...)

3. OTHER MODELS?

TYPES OF STATISTICAL ANALYSES AND MODELS

Based on the research question, we can perform:

- Descriptive
- Inferential
- Predictive
- Causal
- Etc.…

Depending on the type of **Dependent** variable are the models we can use:

- Linear Model: Continuous
- Logit Model: Dichotomous
- Probit Model: Ordinal data
- Multinomial Model: Categorical Data
- Poisson/Binomial: Counting data

The relation between the **Dependent** and the **Independent** dictates the model structure:

- Nested model
- Bayesian

TYPES OF STATISTICAL ANALYSES AND MODELS

Based on
question,

- Descripti
- Inferential
- Predictive
- Causal
- Etc.…

- Linear Model: Continuous
- Logit Model: Dichotomous
- Probit Model: Ordinal data
- Multinomial Model: Categorical Data
- Poisson/Binomial: Counting data

- Nested model
- Bayesian

TYPES OF STATISTICAL ANALYSES AND MODELS

Based on the question, we

- Descriptive
- Inferential
- Predictive
- Causal
- Etc.…

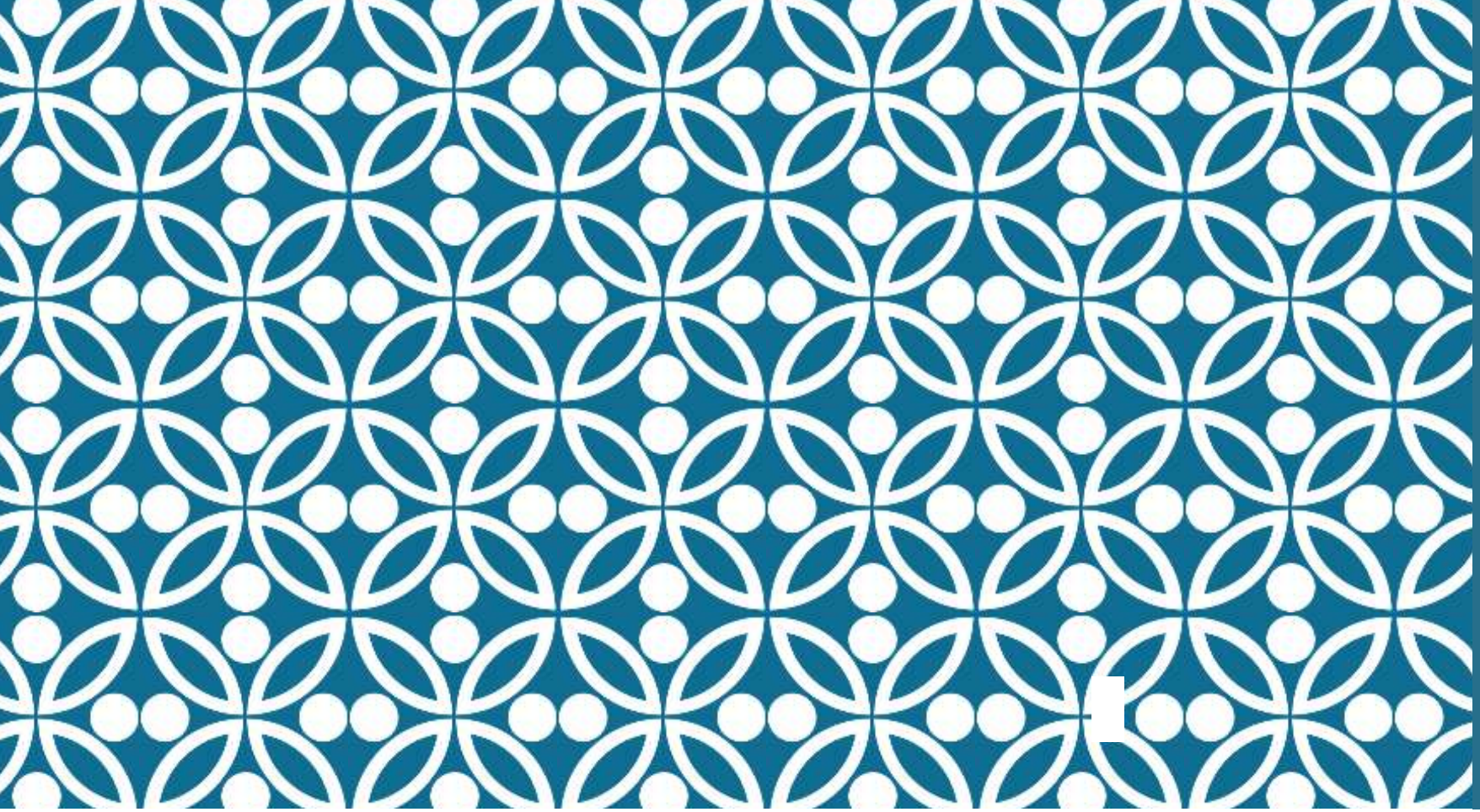
has them all!

They all work with the functions that we learned today. Sometimes, it is just a matter of adjusting some parameters.

- Multinomial Model: Categorical Data
- Poisson/Binomial: Counting data

Using all you new knowledge, try to statistically analyze the Titanic data in R.

EXERCISE 5.1: ANALYZING THE TITANIC DATA



https://sofiag1l.github.io/

THANKS!

Dr. Sofia Gil-Clavel