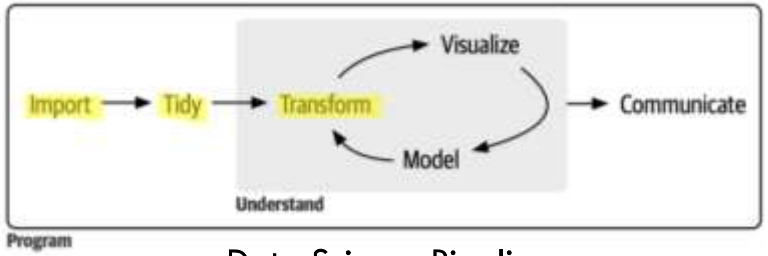
SESSION 4: STATISTICS

DR. SOFIA GIL-CLAVEL

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

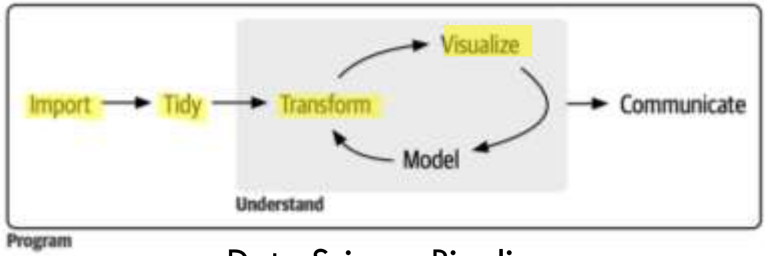
1. SESSION 1-2: RECAP

SESSION 1-2: DATA MANAGEMENT



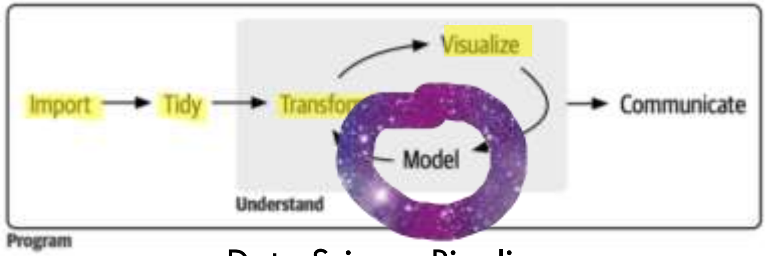
Data Science Pipeline

SESSION 3: DATA VISUALIZATION

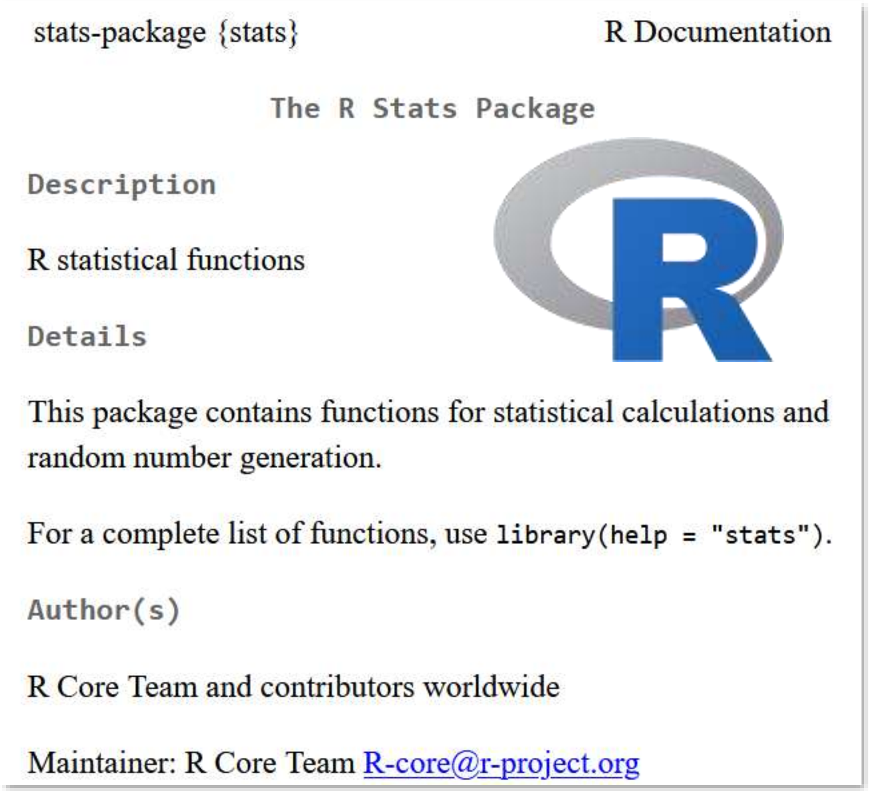


Data Science Pipeline

SESSION 4: 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.

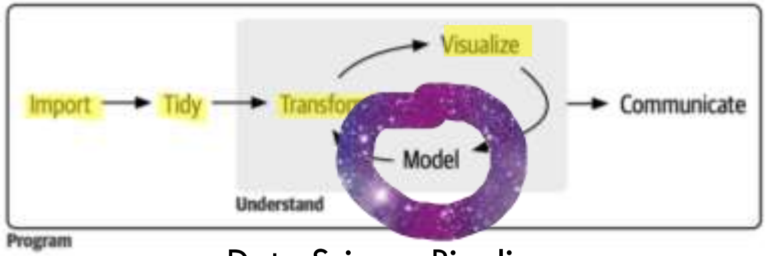
For a complete list of functions, use `library(help = "stats")`.

Author(s)

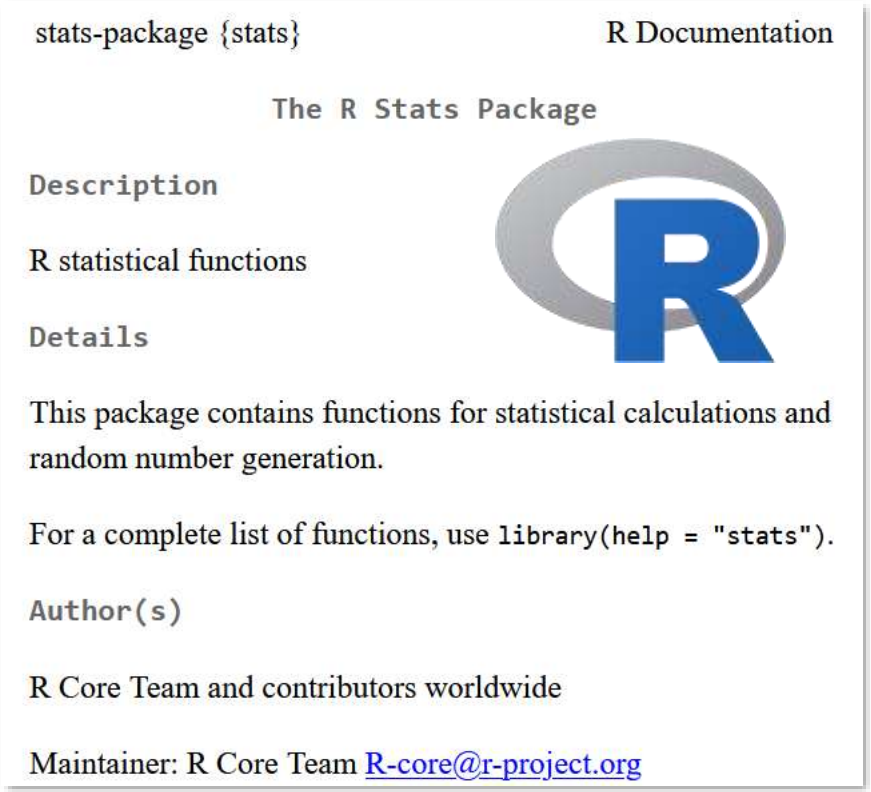
R Core Team and contributors worldwide

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

SESSION 4: 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.

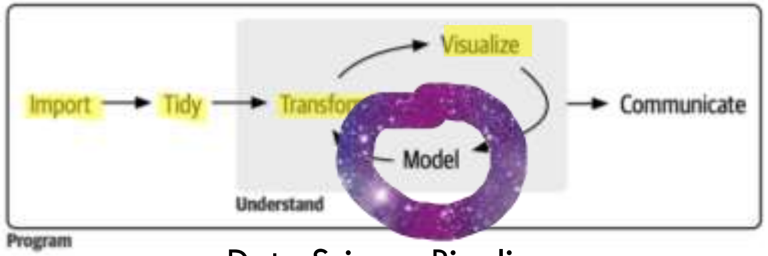
For a complete list of functions, use `library(help = "stats")`.

Author(s)

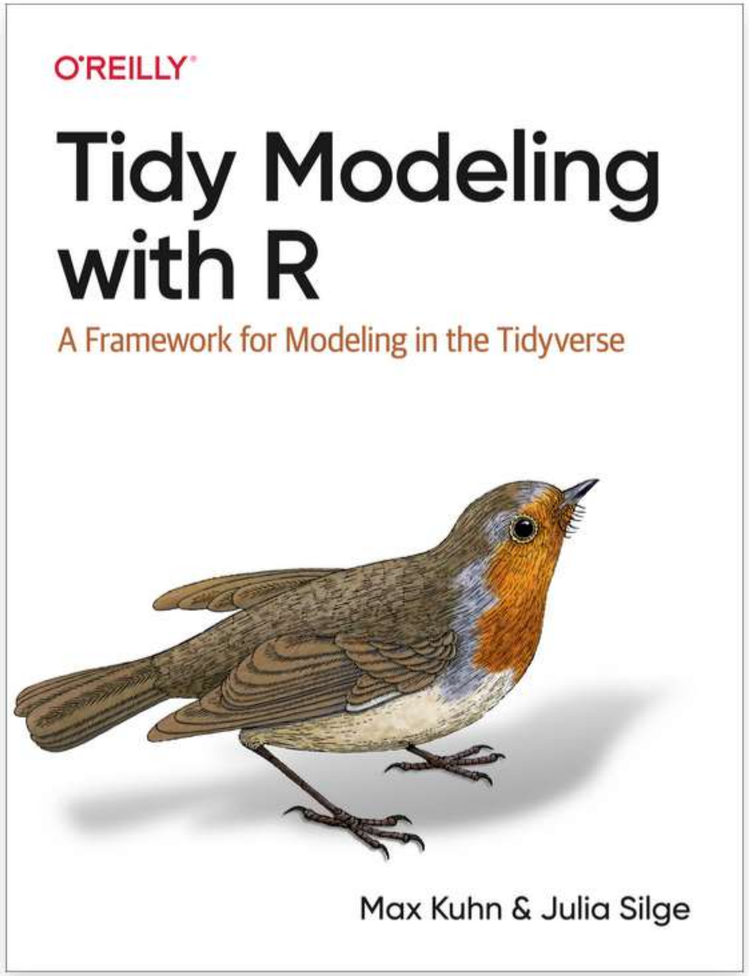
R Core Team and contributors worldwide

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

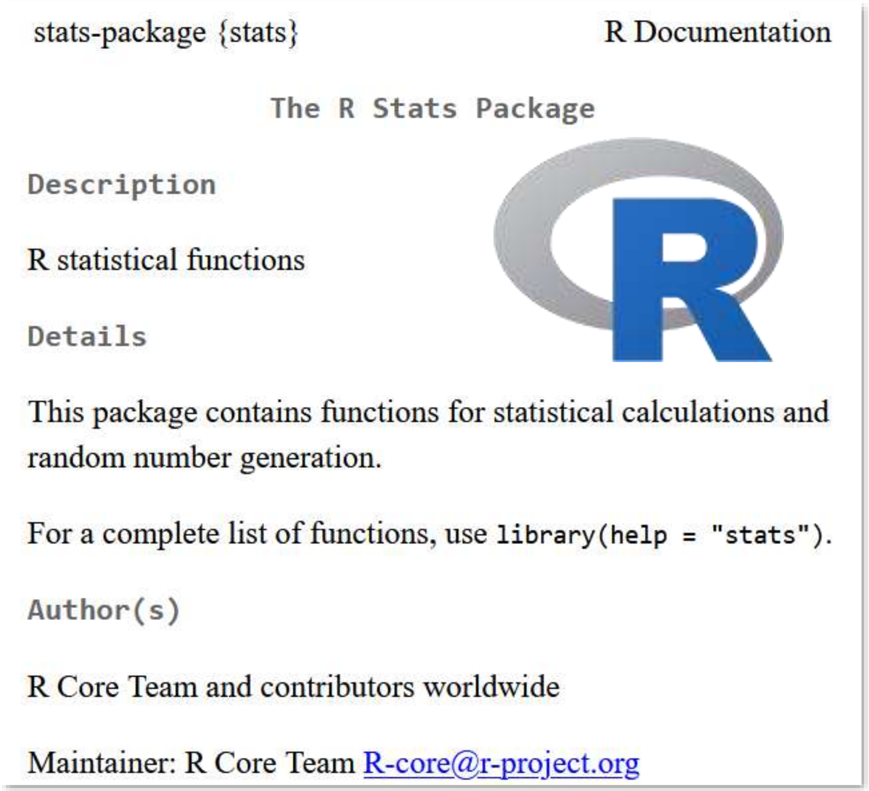
SESSION 4: STATISTICS



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.

$$\boldsymbol{Y} \sim$$

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.

$$\boldsymbol{Y} \sim \boldsymbol{X1} + \boldsymbol{X2} + \boldsymbol{X3}$$

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

$$\boldsymbol{Y} \sim \boldsymbol{X1} + \boldsymbol{X2} + \boldsymbol{X3}$$

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

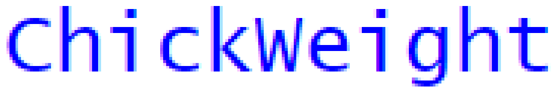
$$\underbrace{Y \sim \boldsymbol{X1} + \boldsymbol{X2} + \boldsymbol{X3}}$$

R was built to understand symbolic language!

BUT FIRST, LET'S OPEN SOME DATA

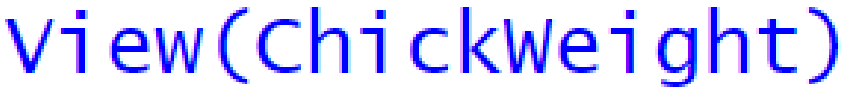
For this section, we will use the database "**ChickWeight**" from the basic R package.

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



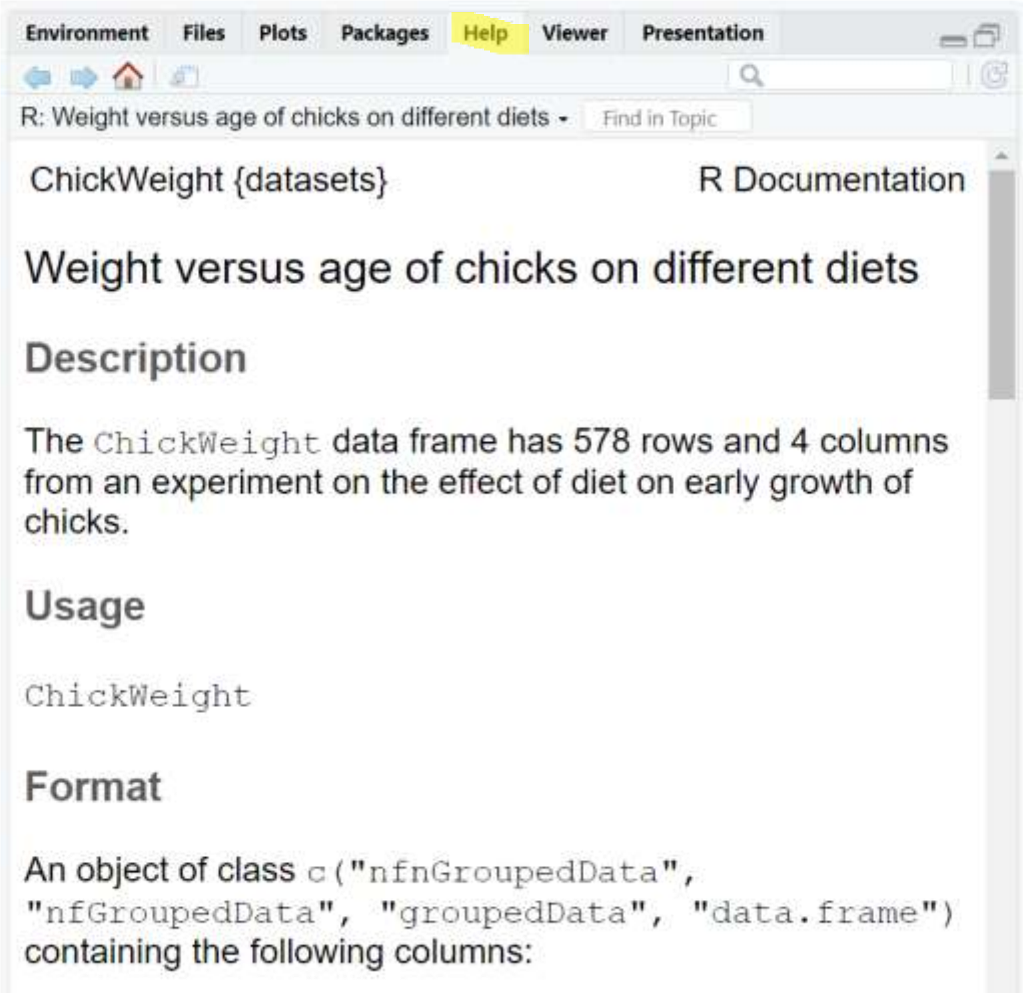
```
ChickWeight
```

To see in an independent window, you can use:



```
View(ChickWeight)
```

BUT FIRST, LET'S OPEN SOME DATA



ChickWeight {datasets} R Documentation

Weight versus age of chicks on different diets

Description

The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.

Usage

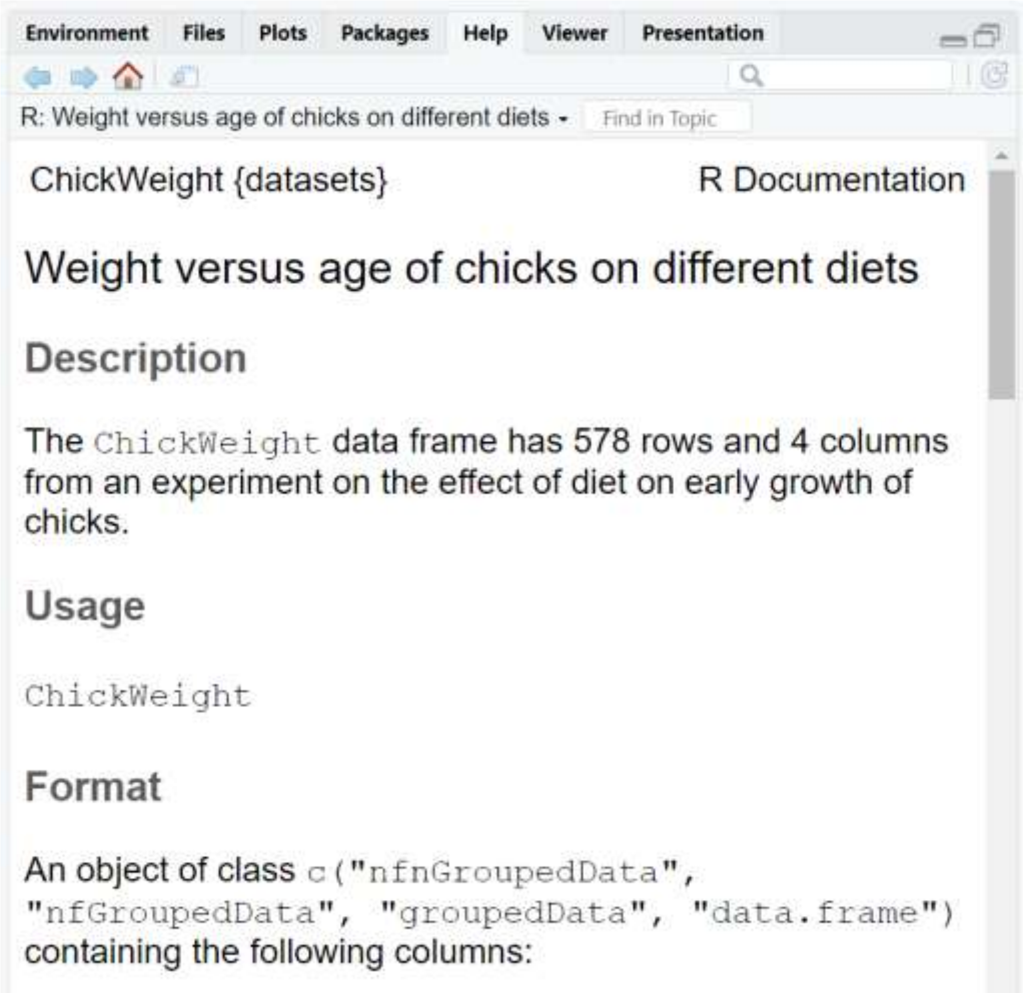
```
ChickWeight
```

Format

An object of class c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame") containing the following columns:

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

BUT FIRST, LET'S OPEN SOME DATA



ChickWeight {datasets} R Documentation

Weight versus age of chicks on different diets

Description

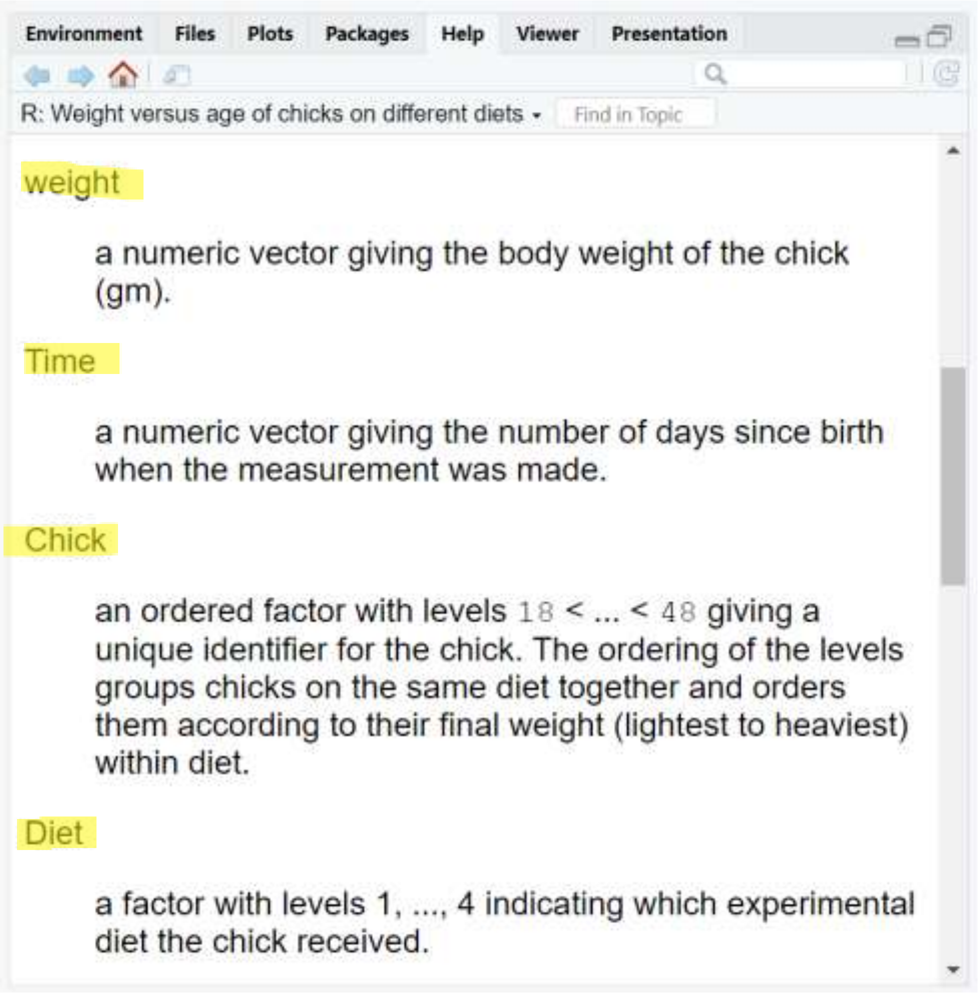
The `ChickWeight` data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.

Usage

```
ChickWeight
```

Format

An object of class `c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame")` containing the following columns:



weight

a numeric vector giving the body weight of the chick (gm).

Time

a numeric vector giving the number of days since birth when the measurement was made.

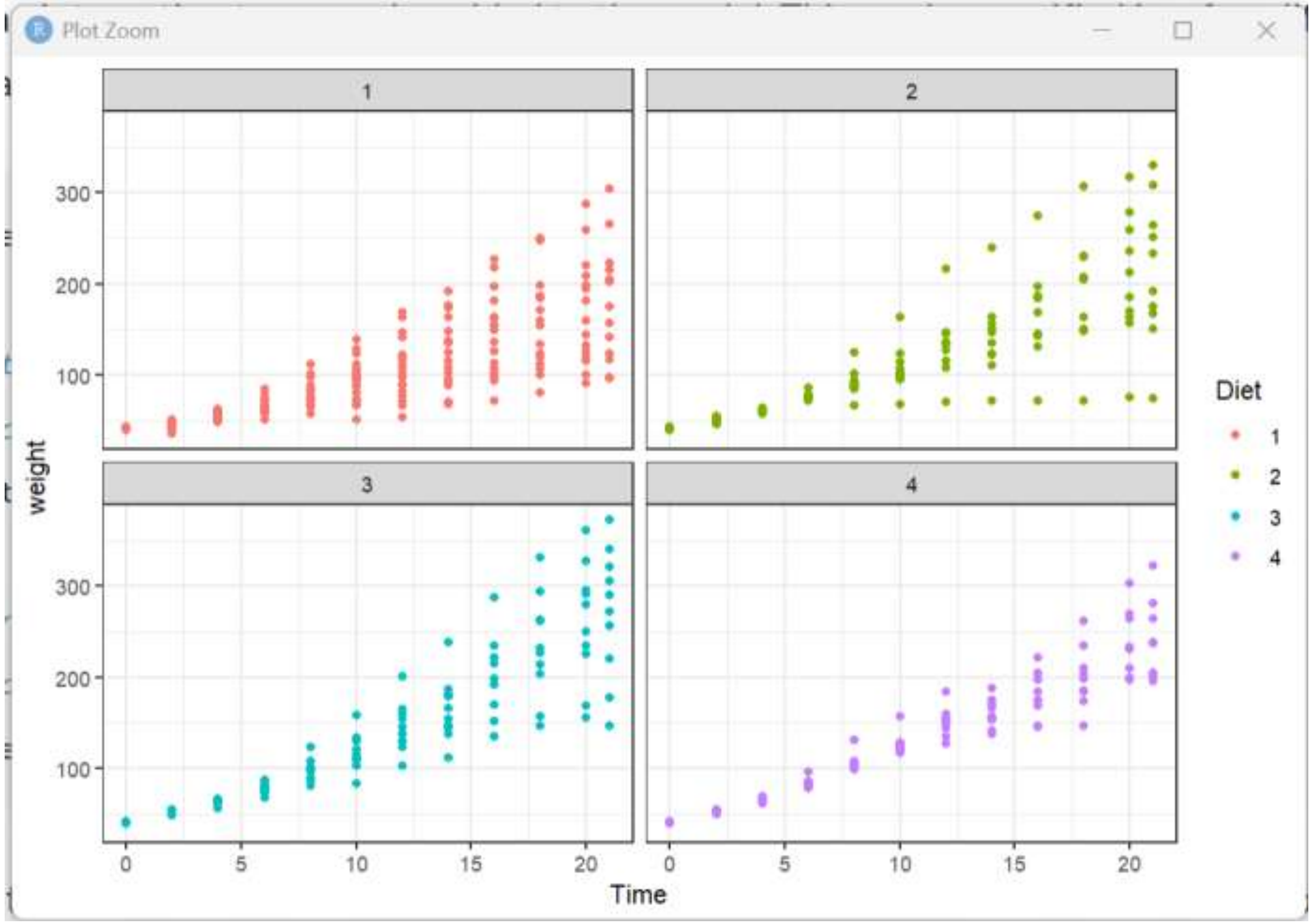
Chick

an ordered factor with levels `18` < ... < `48` giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

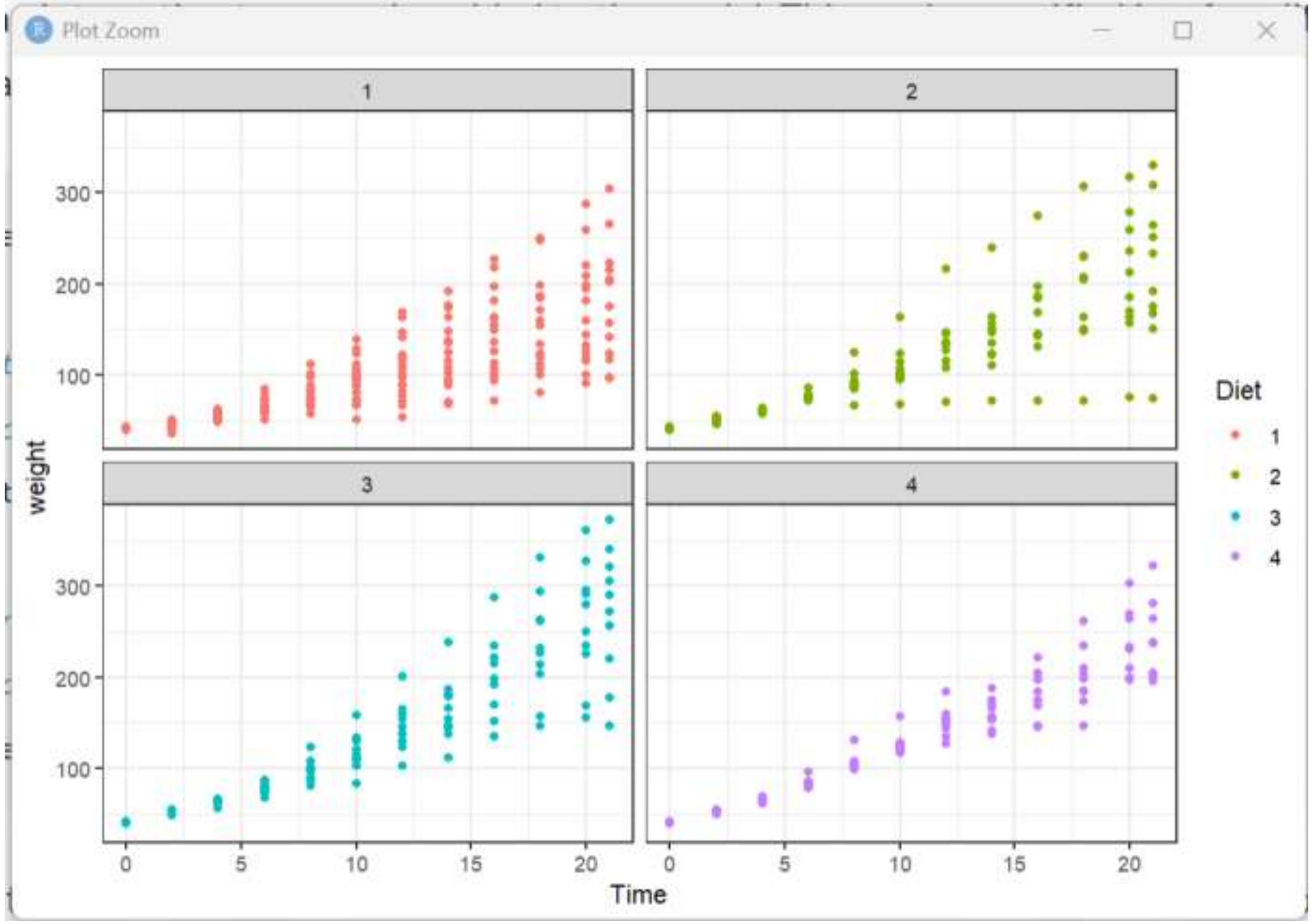
Diet

a factor with levels 1, ..., 4 indicating which experimental diet the chick received.

LET'S CHECK HOW THE DATA LOOKS

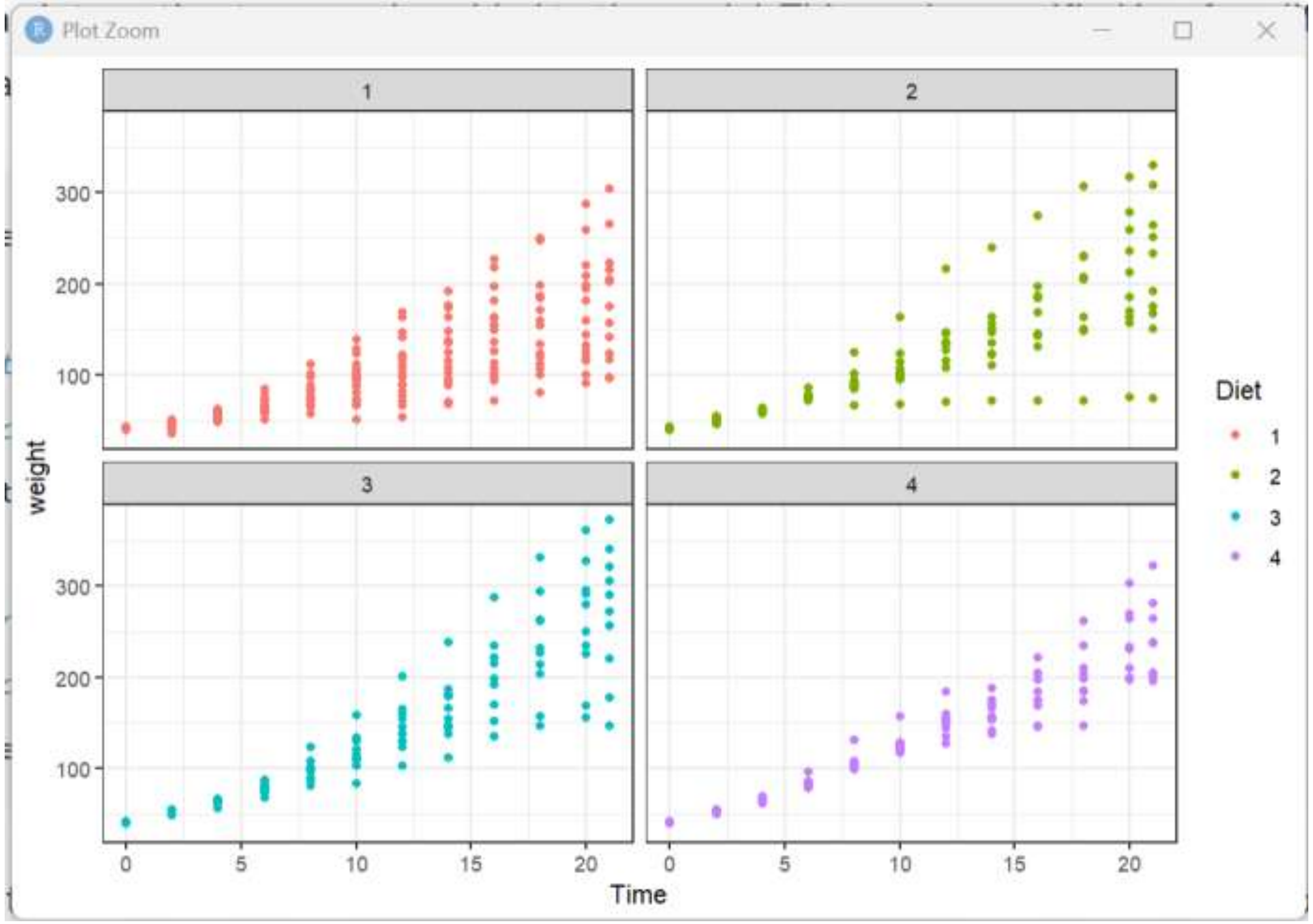


LET'S CHECK HOW THE DATA LOOKS



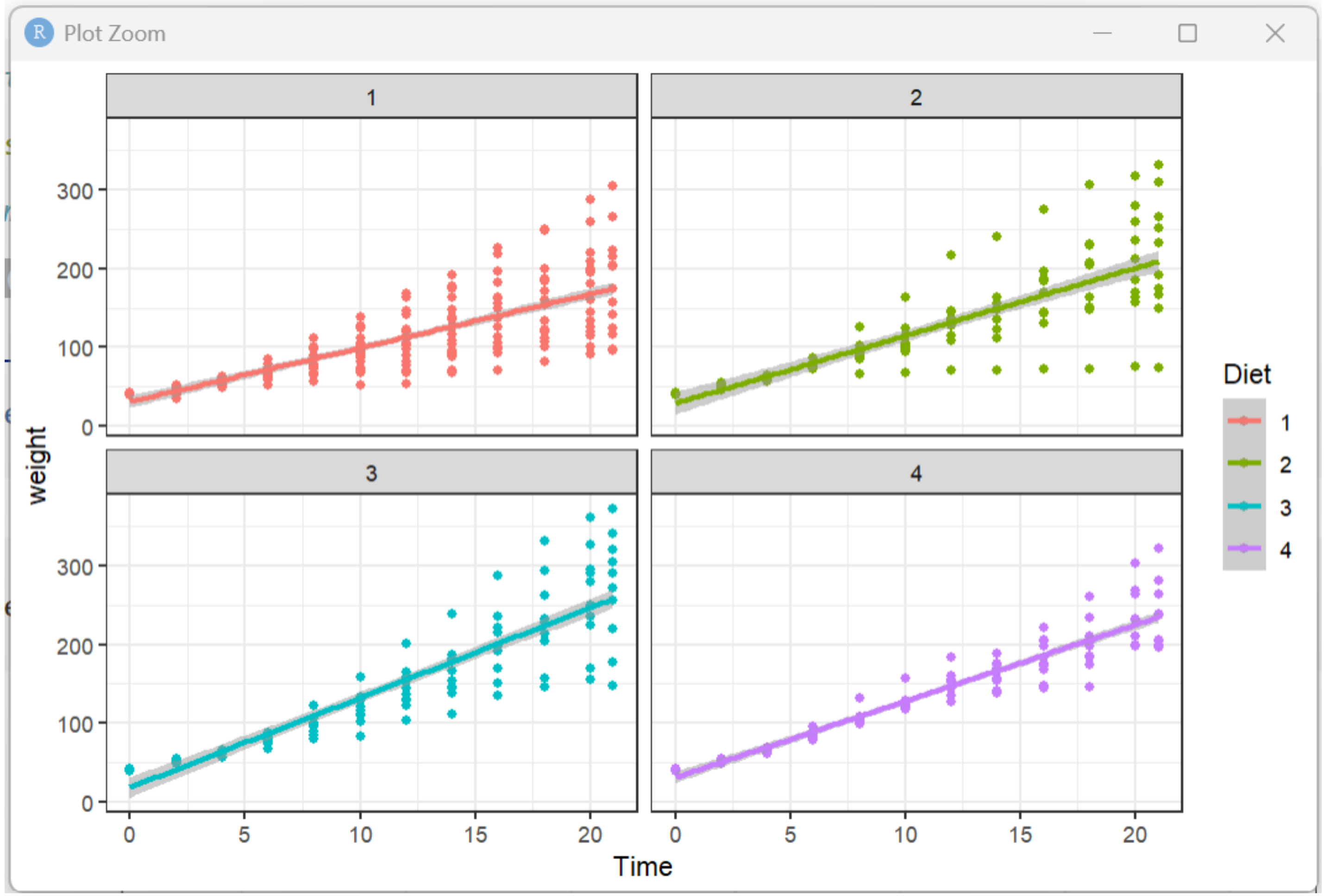
What are some interesting statistical questions?

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

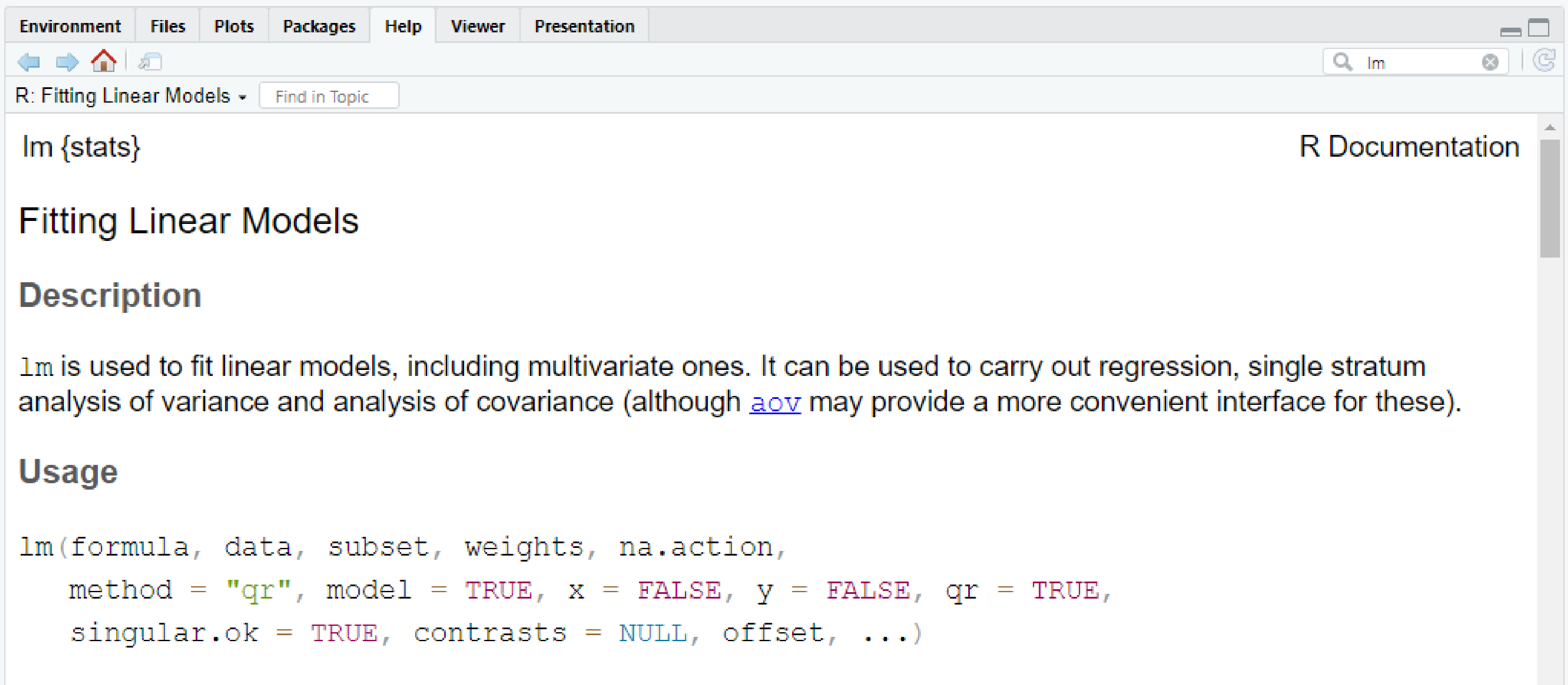


What are some interesting statistical questions? That is up to you!

Now we will learn about:

- Linear regression models

FITTING LINEAR MODELS



lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models, including multivariate ones. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

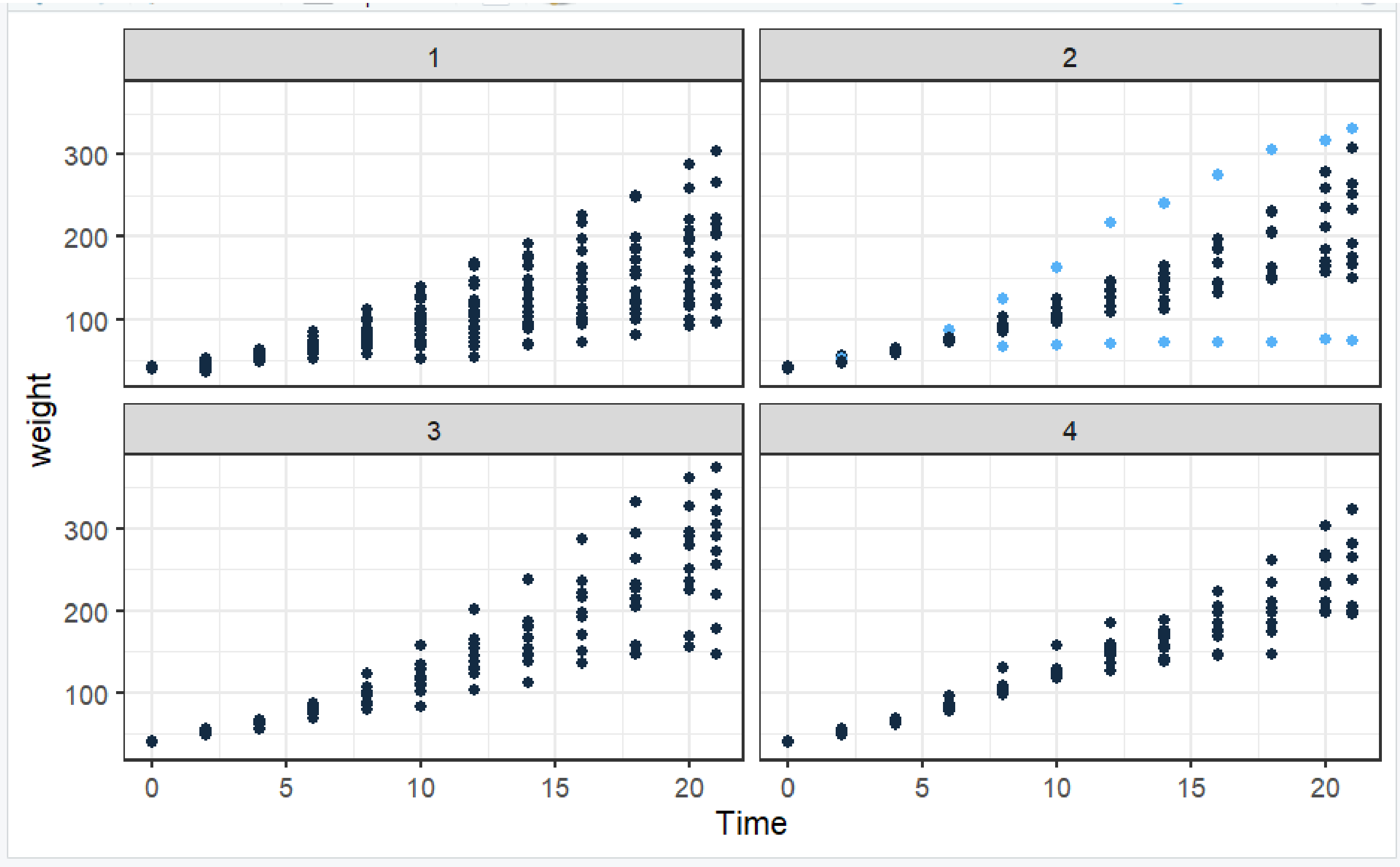
Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

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.



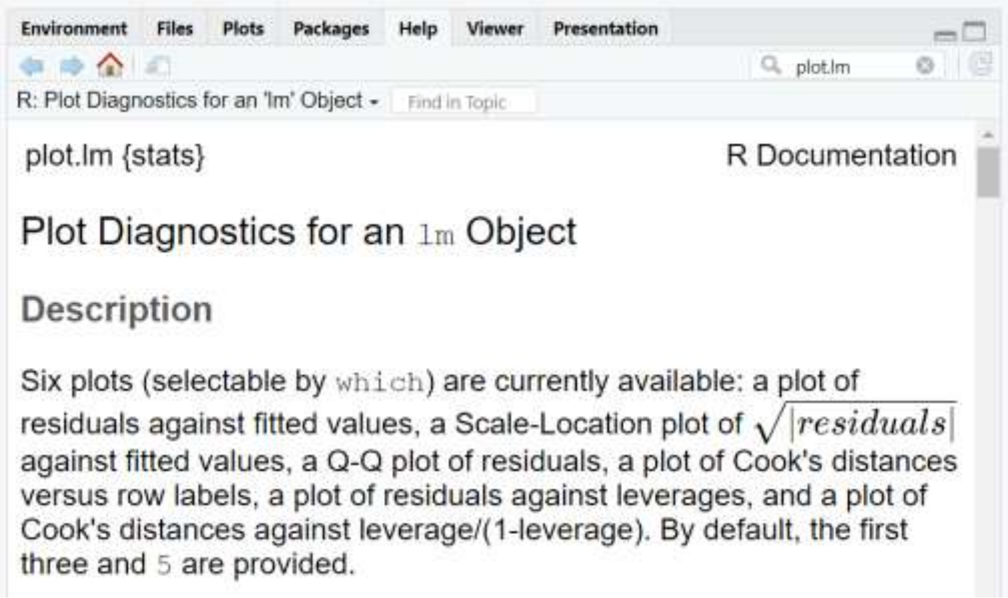
CHECKING THE ASSUMPTIONS

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



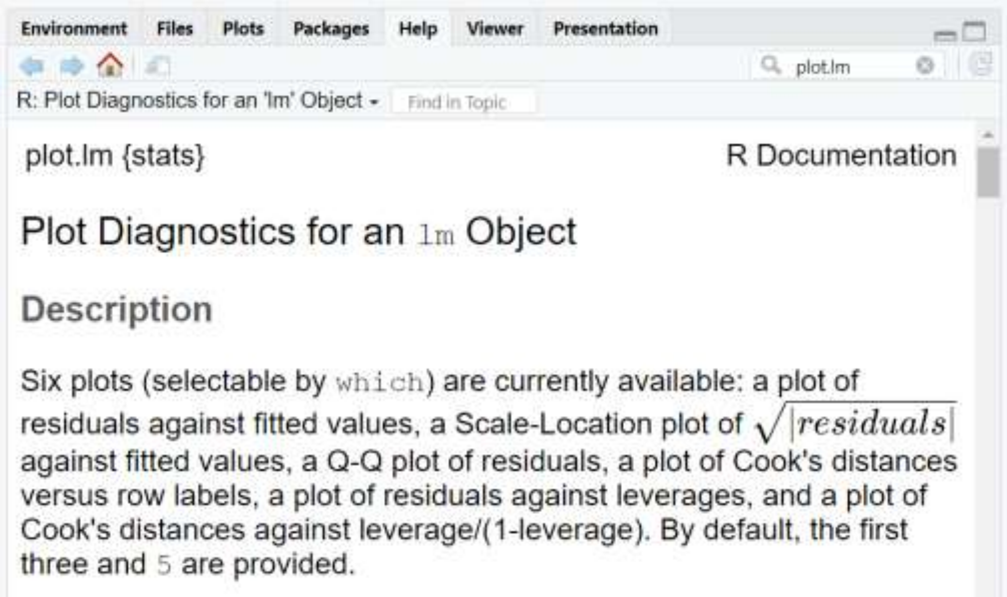
plot.lm {stats} R Documentation

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.

CHECKING THE ASSUMPTIONS

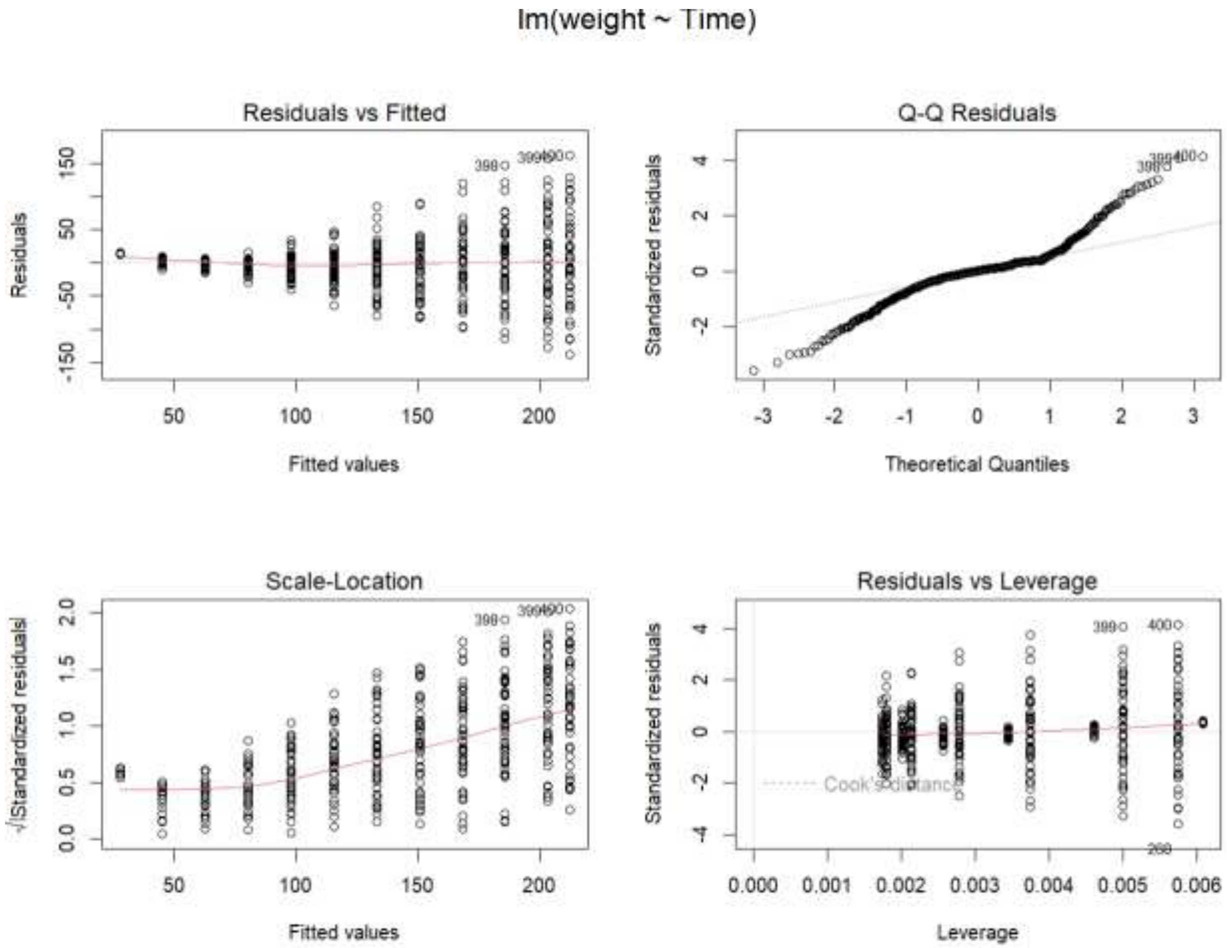


plot.lm {stats} R Documentation

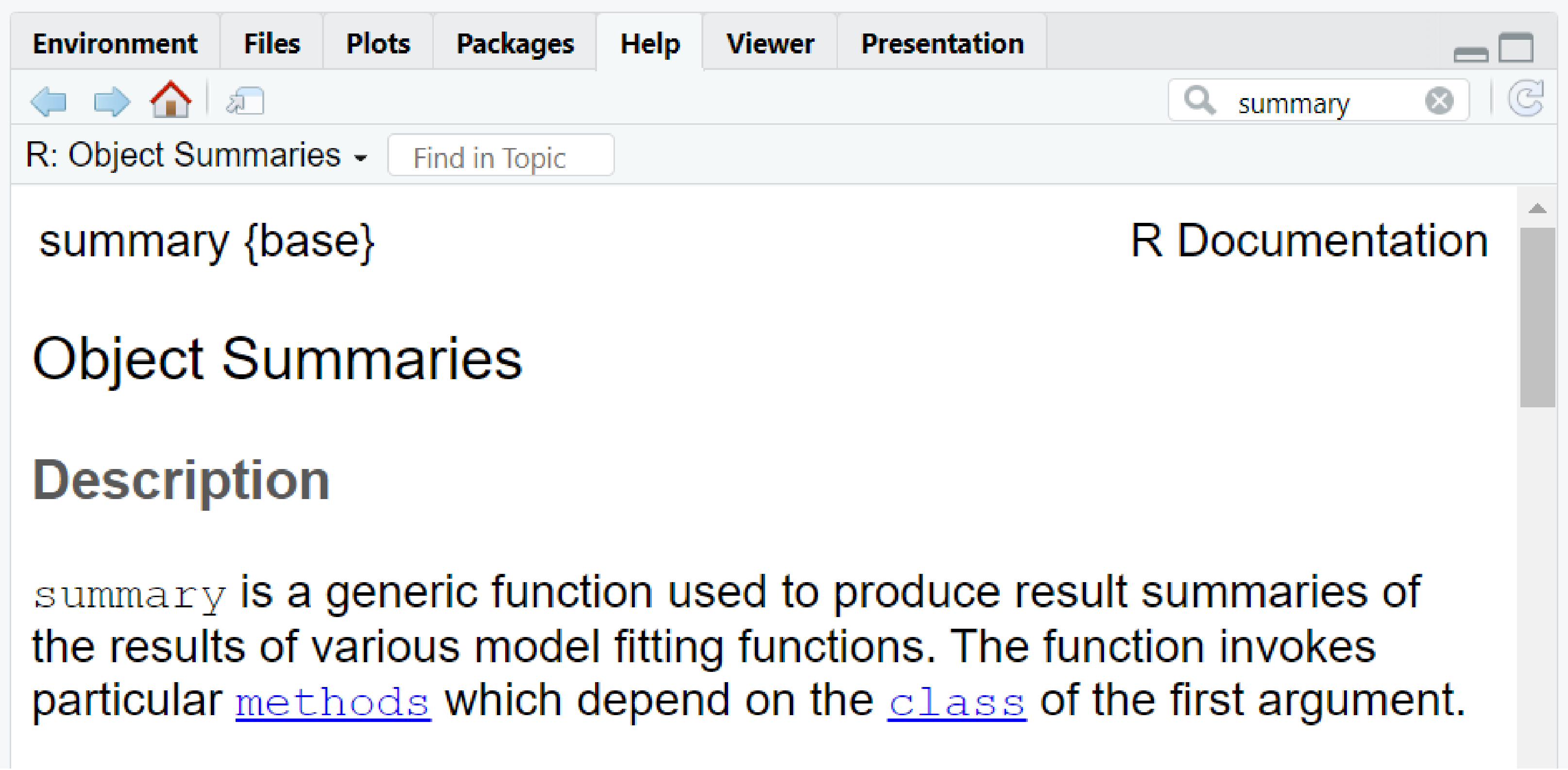
Plot Diagnostics for an `lm` Object

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



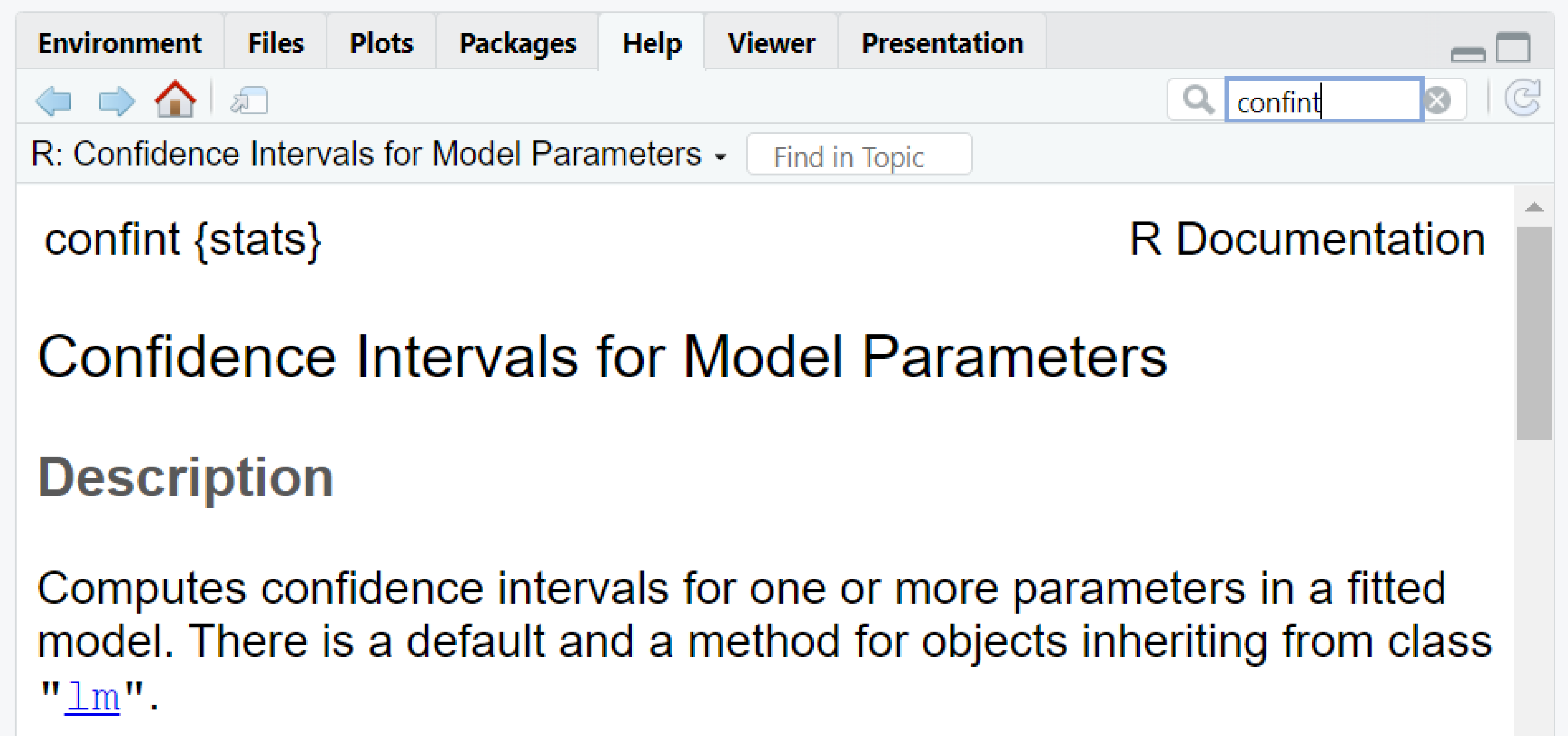
summary {base}   R Documentation

Object Summaries

Description

`summary` is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular `methods` which depend on the `class` of the first argument.

CONFIDENCE INTERVALS



R: Confidence Intervals for Model Parameters

confint {stats} R Documentation

Confidence Intervals for Model Parameters

Description

Computes confidence intervals for one or more parameters in a fitted model. There is a default and a method for objects inheriting from class "lm".

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).

$$\boldsymbol{Y \sim X1 \times X2}$$

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

$$\boldsymbol{weight \sim Time \times Diet}$$

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).

$$\boldsymbol{Y \sim X1 \times X2}$$

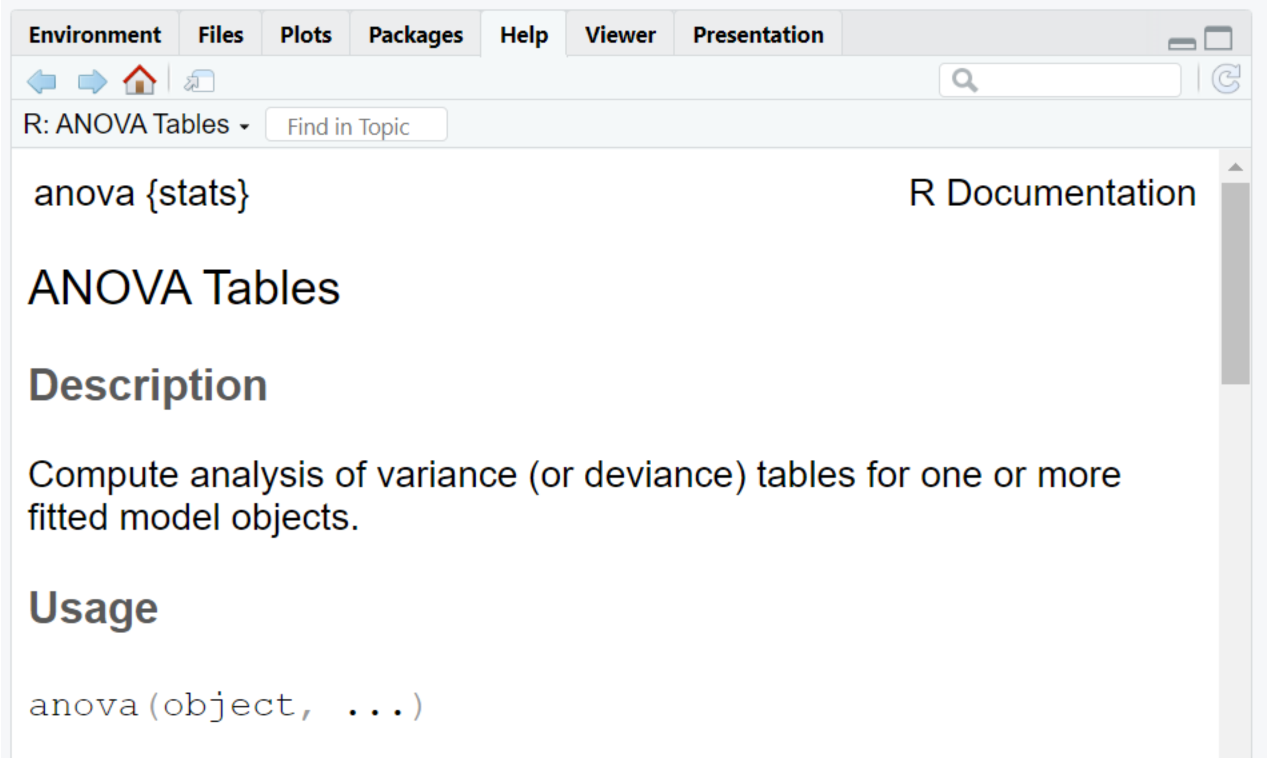
So, in our data, we could think that **the effect that time has on the chickens' weight depends on the diet:**

$$\boldsymbol{weight \sim Time \times Diet}$$

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 {stats} R Documentation

ANOVA Tables

Description

Compute analysis of variance (or deviance) tables for one or more fitted model objects.

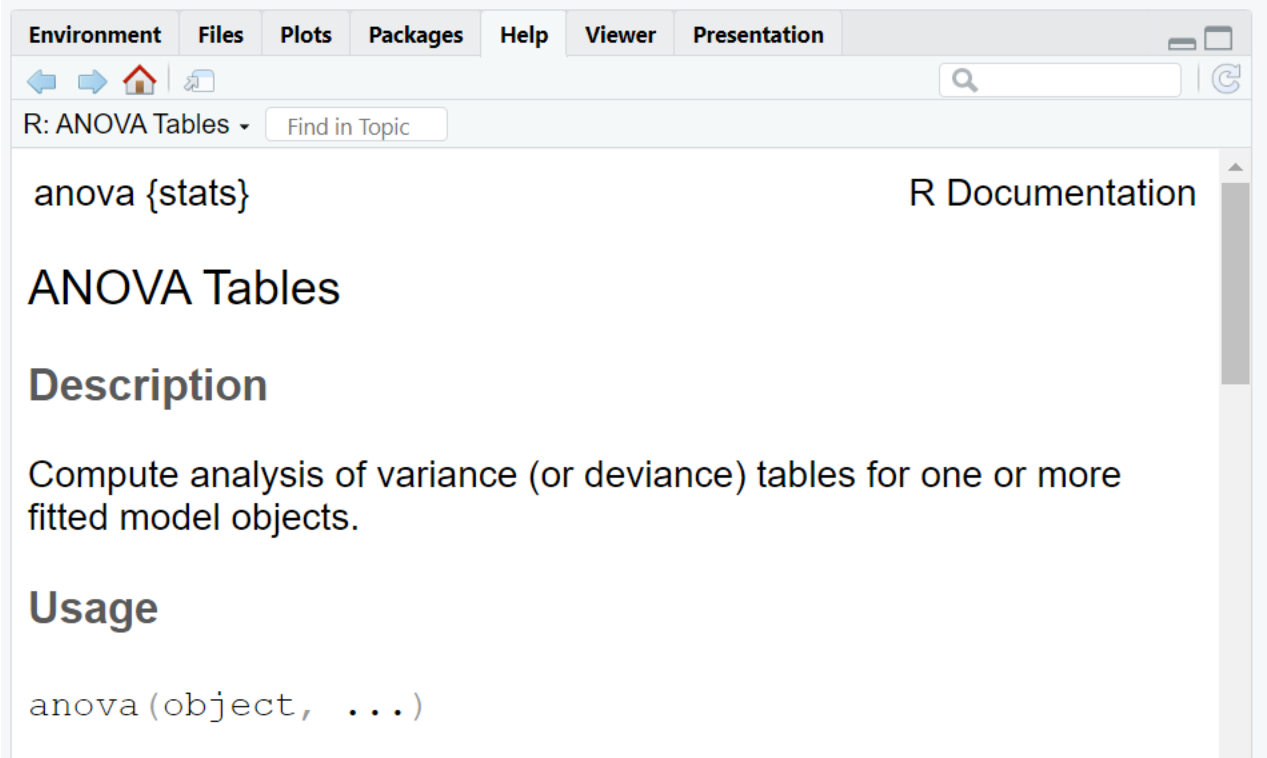
Usage

```
anova(object, ...)
```

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?



anova {stats} R Documentation

ANOVA Tables

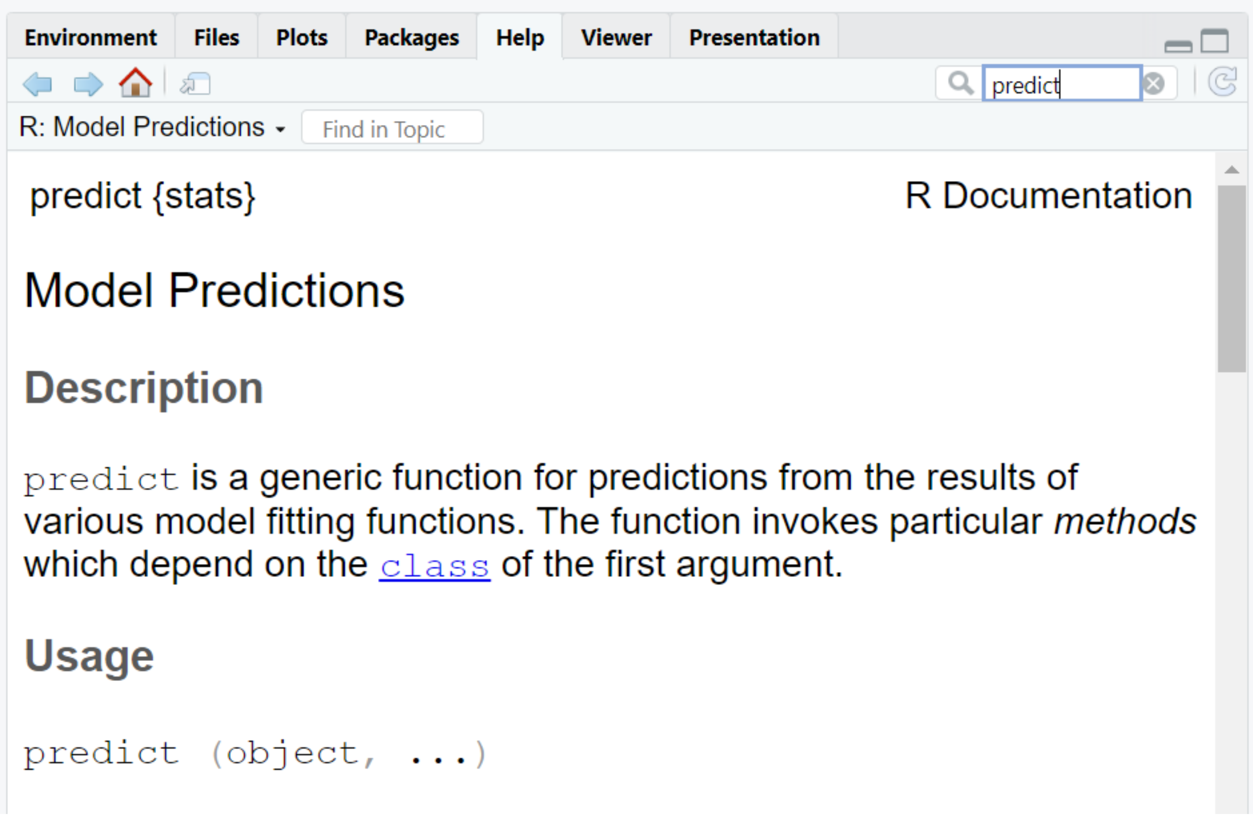
Description

Compute analysis of variance (or deviance) tables for one or more fitted model objects.

Usage

```
anova(object, ...)
```

PREDICTING



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

R has them all!

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

e

ne

- Nested model
- Bayesian

TYPES OF STATISTICAL ANALYSES AND MODELS

has them all!

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

Based on the question, we

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

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

n the

s the

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

EXERCISE 3.1: ANALYZING THE TITANIC DATA

Join us!

https://forms.office.com/e/8Bgd2YsasJ

All about the lab:

https://societal-analytics.nl/

Contact us at:

analytics-lab.fsw@vu.nl.

https://sofiag1l.github.io/

THANKS!

Dr. Sofia Gil-Clavel