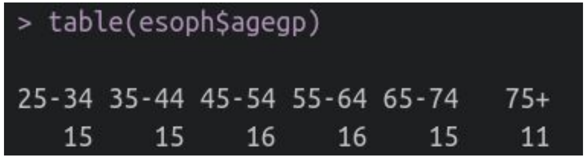
Análise de dados em R

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Summarize

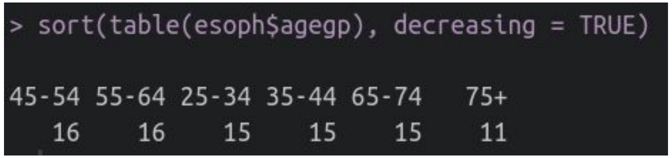
- Frequency table: frequency of each value



```
> table(esoph$agegp)

25-34 35-44 45-54 55-64 65-74   75+ 
   15    15    16    16    15    11 
```

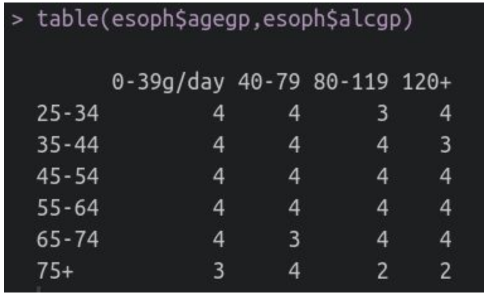
- Mode: the most frequent value



```
> sort(table(esoph$agegp), decreasing = TRUE)

45-54 55-64 25-34 35-44 65-74   75+ 
   16    16    15    15    15    11 
```

- Contingency tables: cross-frequency of values for two variables



```
> table(esoph$agegp,esoph$alcgp)
       
        0-39g/day 40-79 80-119 120+
  25-34         4     4      3    4
  35-44         4     4      4    3
  45-54         4     4      4    4
  55-64         4     4      4    4
  65-74         4     3      4    4
  75+           3     4      2    2
```

Before start!

Ggplot2 - install.package(“ggplot2”)
Esquisse - install.package(“esquisse”)

Data Explorer - install.package(“DataExplorer”)

Overview

- **EDA**
- **Visualization**
- **Correlation**
- **Test Hypothesis**

Exploration Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics. We can do this

- **Summarize**
- **Visualize**
- **Correlation**
- **Test Hypothesis**

Summarization

- summary
- group_by
  - mean
  - sd
  - sum
  - quantiles
  - …functions that aggregate

• **Mean (or sample mean) - sensitive to extreme values**

• **Median: It is the 50th-precentile, i.e. the value above (below) which there are 50% of the values in the data set**

• **Mode: It is the most common (more frequently occurring) value in a set of values. Note that the mode can be applied to categorical variables**

- **Variance - sensitive to extreme values**
-

$$\sigma_X^2$$

• **Standard Deviation - sensitive to extreme values**

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)^2}$$

• **Inter-quartile Range (IQR)**

  • **It is the difference between the 3rd (Q3) and 1st (Q1) quartiles**

  • **Q1 is the number below which there are 25% of the values**

  • **Q3 is the number below which there are 75% of the values**

Visualization

Categorical Variables

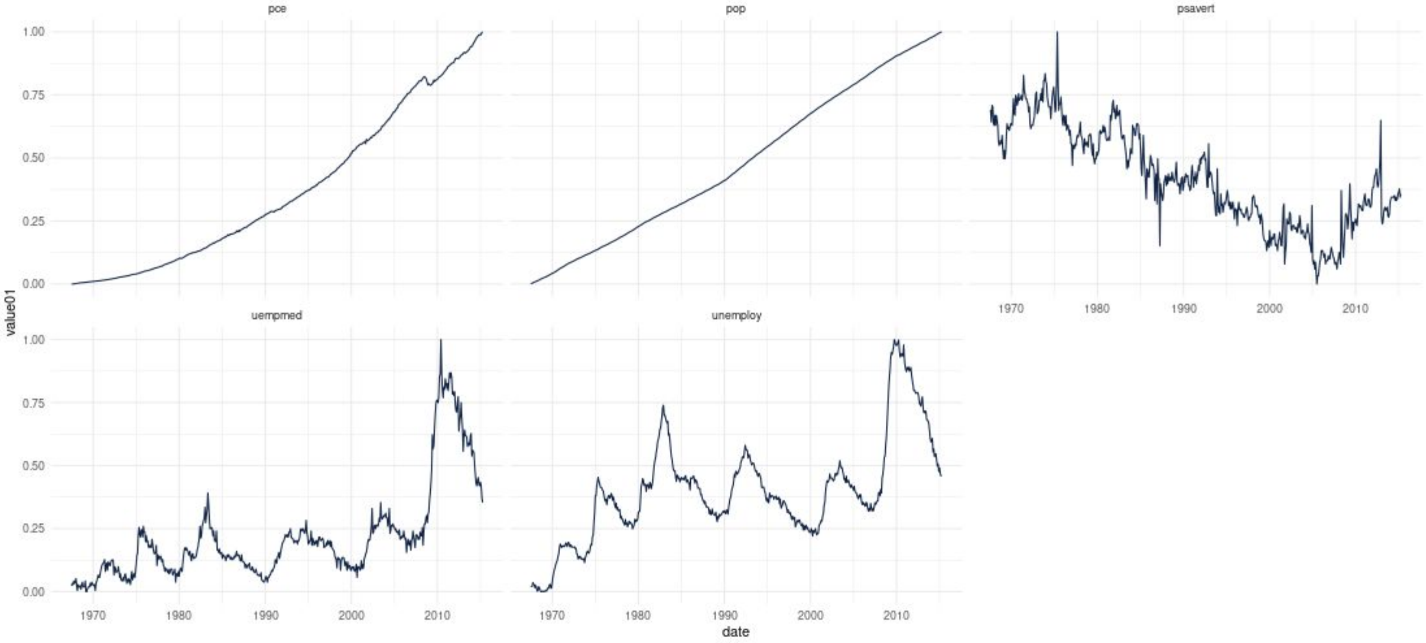
- Barplots
- Piecharts
- . . .

Numeric Variables

- Histograms
- QQ Plots
- Boxplots
- . . .

Visualization

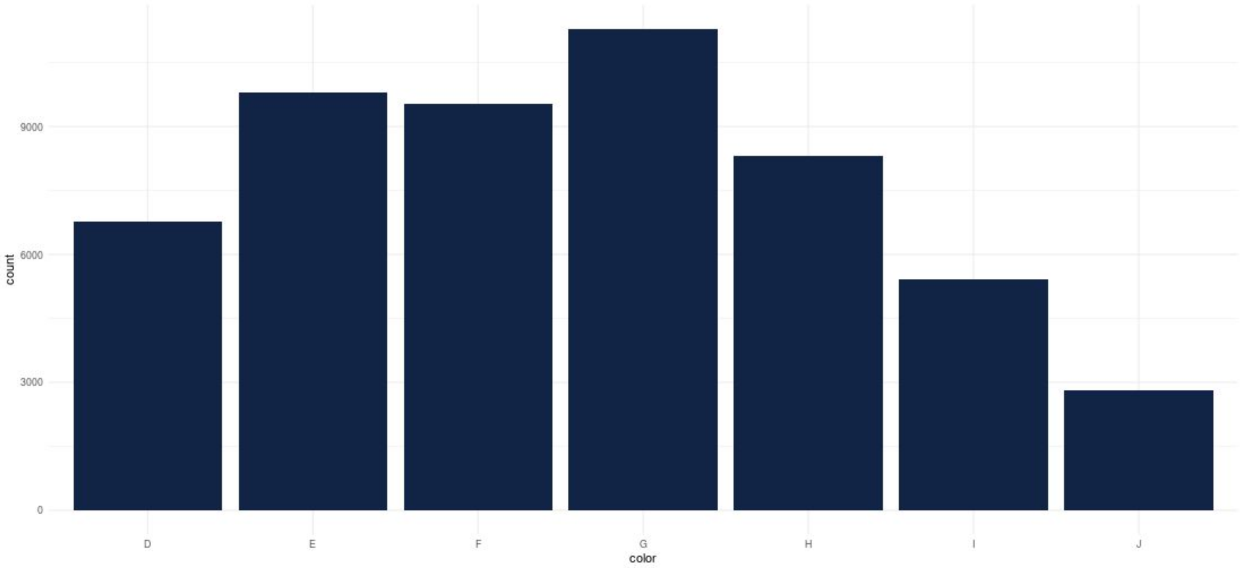
Line plot is a chart which displays information as a series of data points called 'markers' connected by straight line segments.



Visualization

Barplots

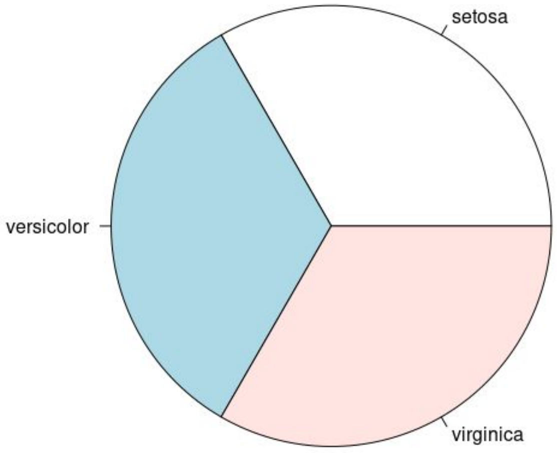
- The main purpose is to display a set of values as heights of bars
- It can be used to display the frequency of occurrence of different values of a categorical variable



Visualization

Piecharts

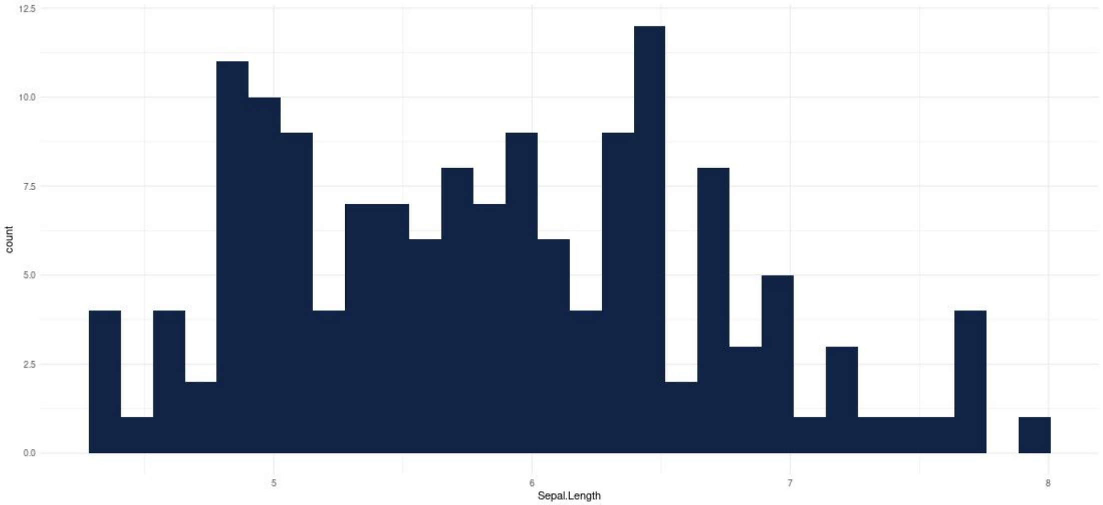
- **Have the same purpose as bar plots but with information in the form of a pie.**
- **Are not so good for comparison purposes**



Visualization

Histograms

- The main purpose is to display how the values of a continuous variable are distributed
- It is obtained as follows:
  - first, the range of the variable is divided into a set of bins (intervals of values)
  - then, the number of occurrences of values on each bin is counted
  - then, this number is displayed as a bar



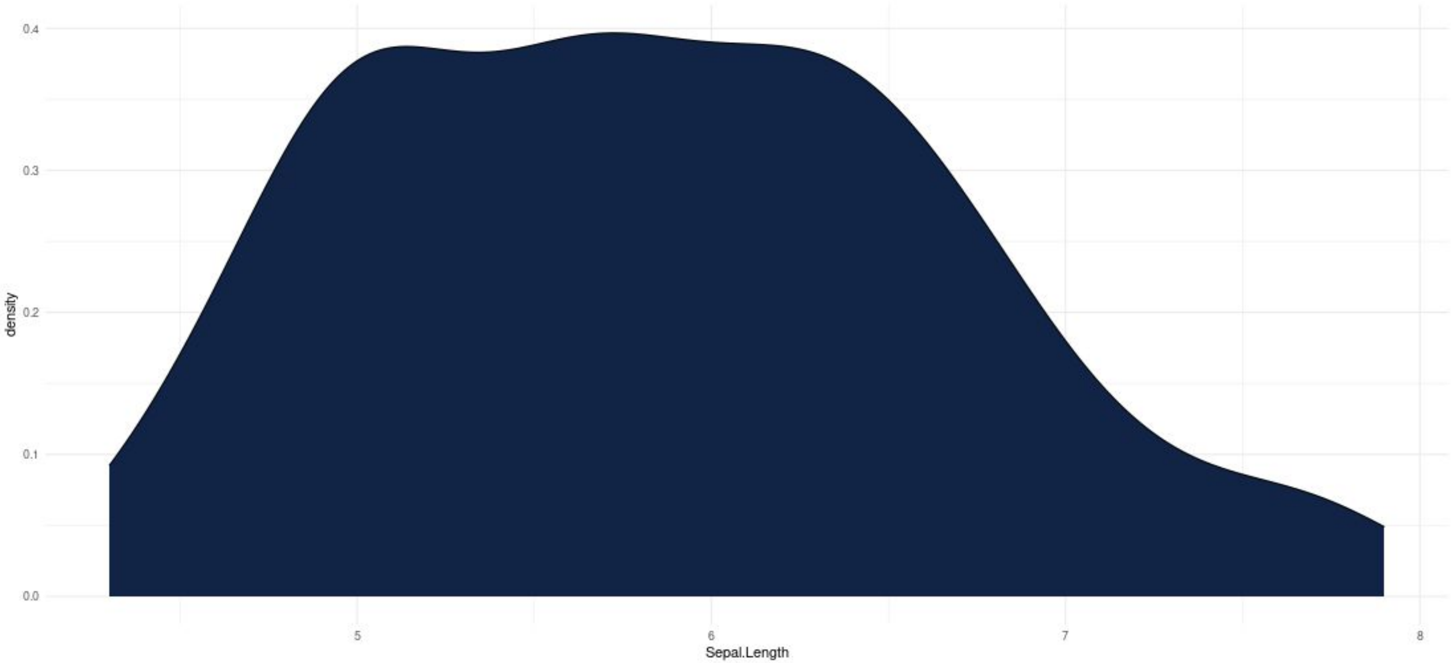
Visualization

Problems with Histograms

- **Histograms may be misleading in small data sets**
- **The shape of the histogram depends on the number of bins**
- **How are the limits of the bins chosen? There are several algorithms for this.**

Visualization

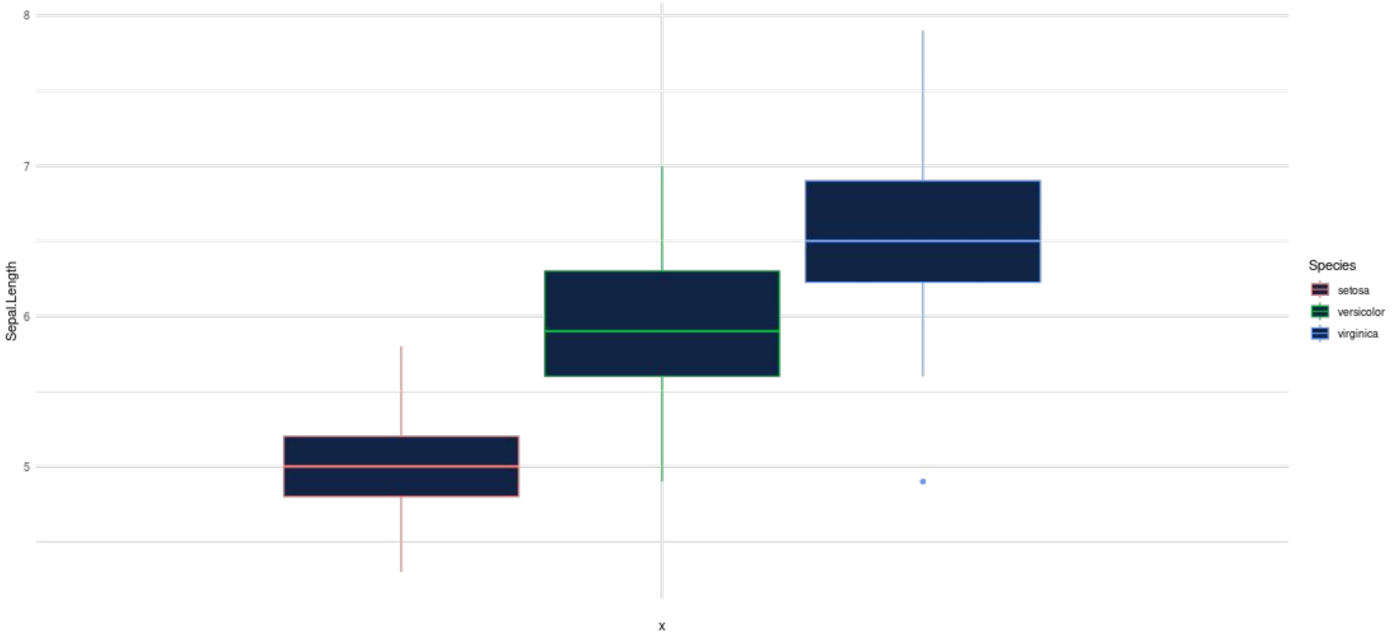
Density plots
Some of the problems of histograms can be tackled by smoothing the estimates of the distribution of the values. That is the purpose of kernel density estimates



Visualization

Boxplots

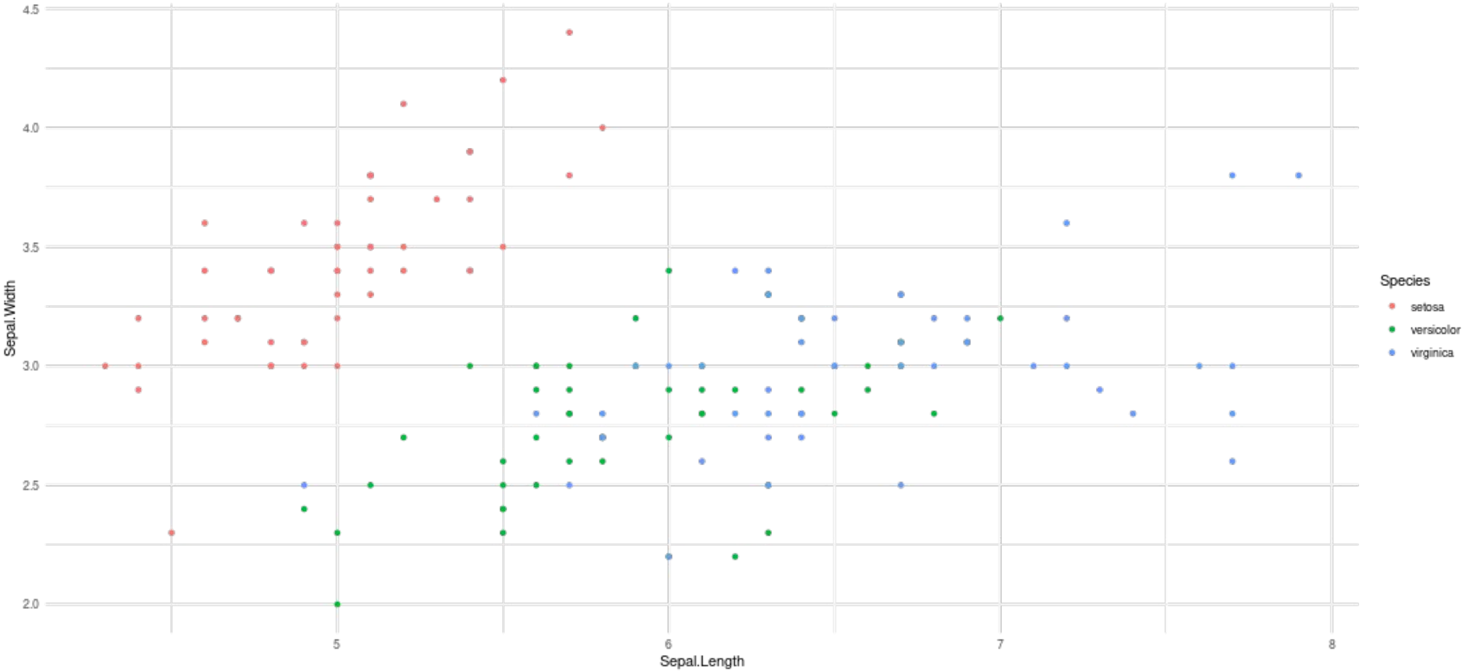
- **Box plot provide an interesting summary of a variable distribution**
- **For instance, they inform us of the interquartile range and of the outliers (if any)**



Visualization

Scatterplots

- The natural graph for showing the relationship between two numeric variables



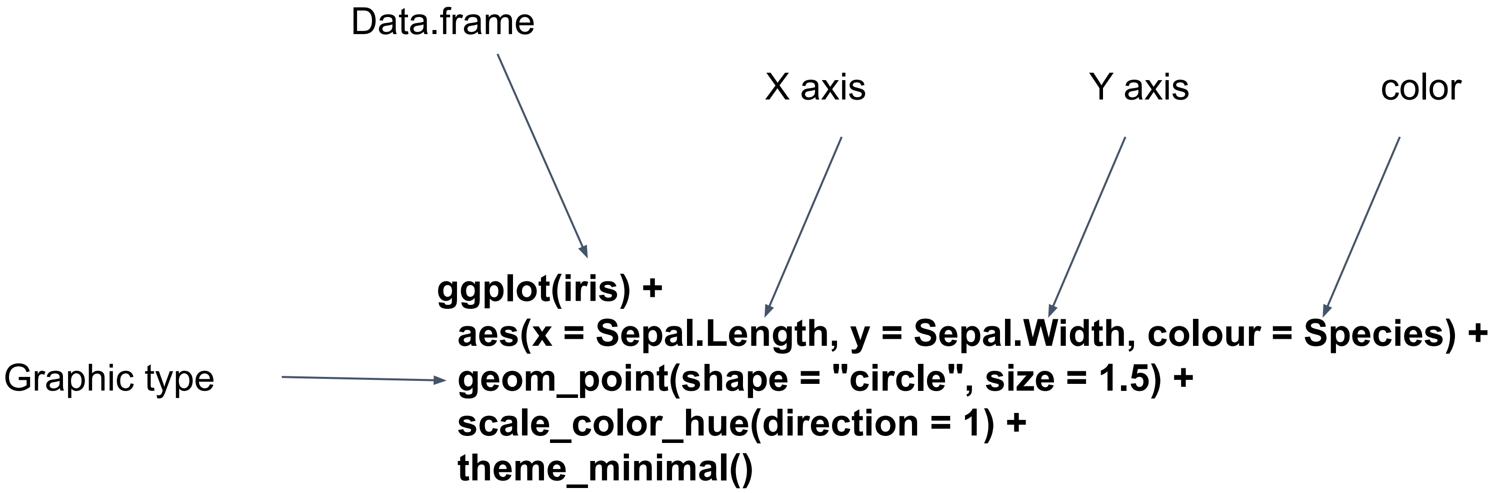
ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Cheat sheet:
https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf

Gallery:
https://r-graph-gallery.com/

ggplot2



Data.frame

X axis

Y axis

color

Graphic type

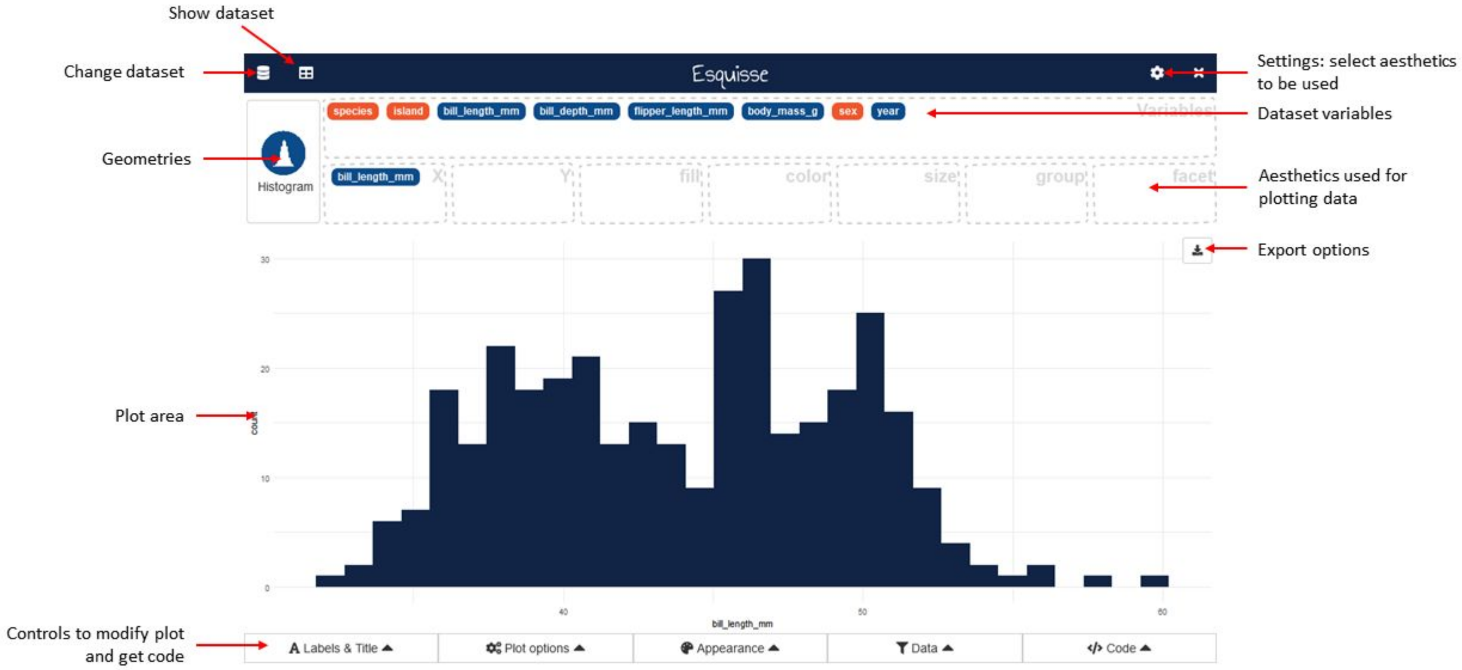
```
ggplot(iris) +
  aes(x = Sepal.Length, y = Sepal.Width, colour = Species) +
  geom_point(shape = "circle", size = 1.5) +
  scale_color_hue(direction = 1) +
  theme_minimal()
```

Esquisse

Esquisse is an R package which creates easy ggplot charts through a drag and drop interface

https://cran.r-project.org/web/packages/esquisse/vignettes/get-started.html

Esquisse



Show dataset

Change dataset

Geometries

Plot area

Controls to modify plot and get code

Settings: select aesthetics to be used

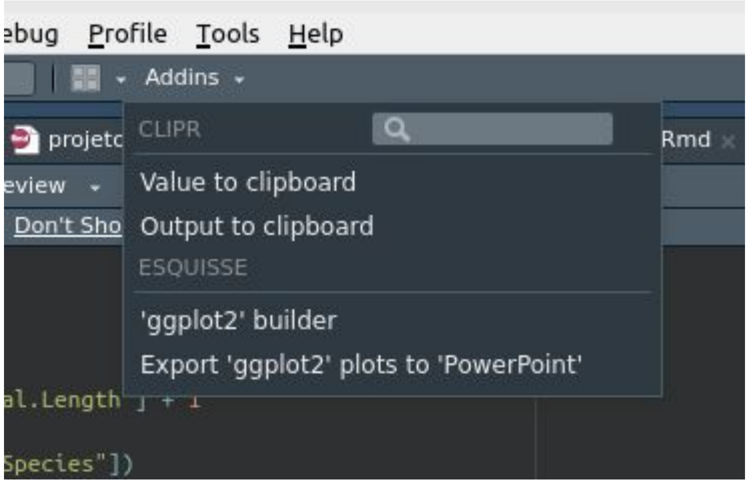
Dataset variables

Aesthetics used for plotting data

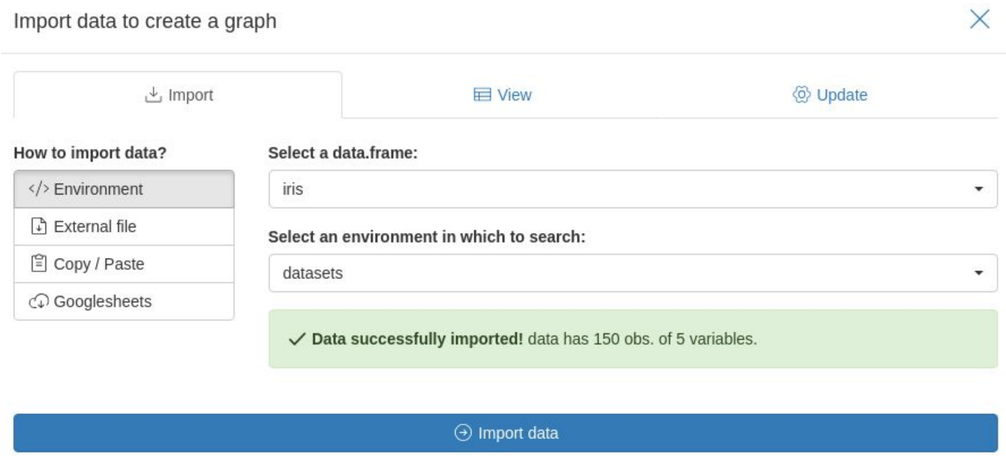
Export options

Esquisse

1) Open Esquisse



2) Select data.frame



Import data to create a graph

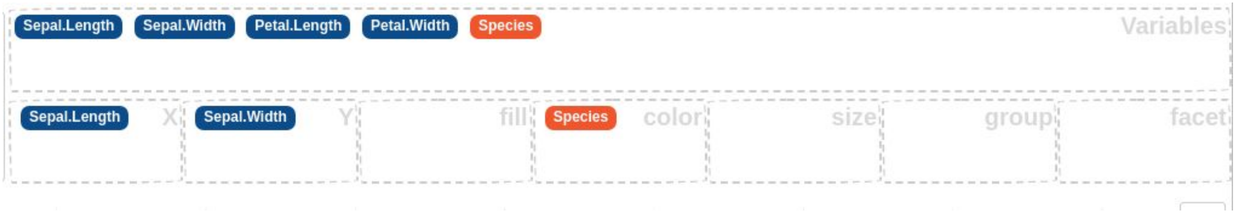
How to import data? Environment

Select a data.frame: iris

Select an environment in which to search: datasets

✓ **Data successfully imported!** data has 150 obs. of 5 variables.

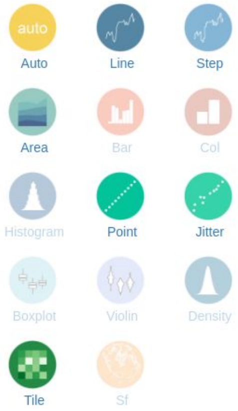
Esquisse



2) Select variables for each plot component

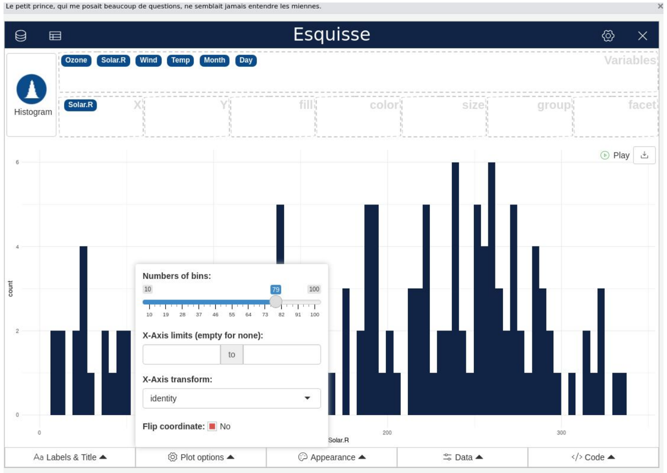
X - X axis
Y - Y axis
fill - fill the area with different color for different values
color - line with different color for different values
size - size of the points increase if value increases
group - group plots each categorical value
facet - create differents plots for each categorical value

Esquisse



3) Plot type

4) See plot and change some parameters



Le petit prince, qui me posait beaucoup de questions, ne semblait jamais entendre les miennes.

Numbers of bins: 79

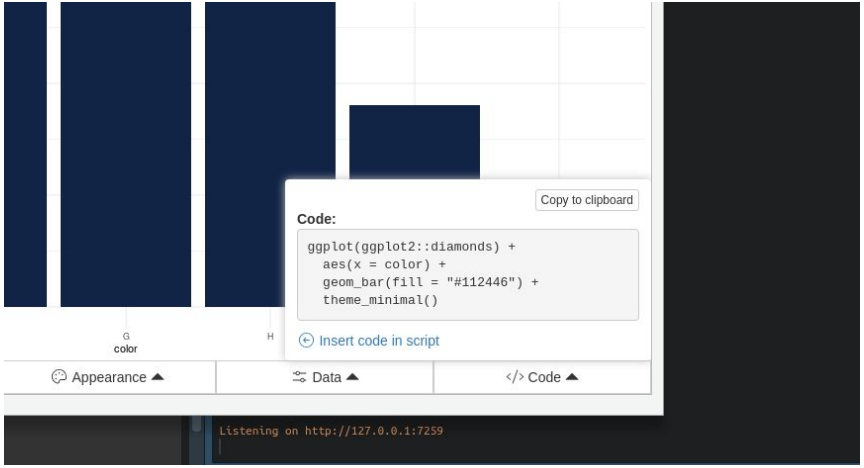
X-Axis limits (empty for none): to

X-Axis transform: identity

Flip coordinate: No

Esquisse

Finally, get the code to produce the plot



Code:

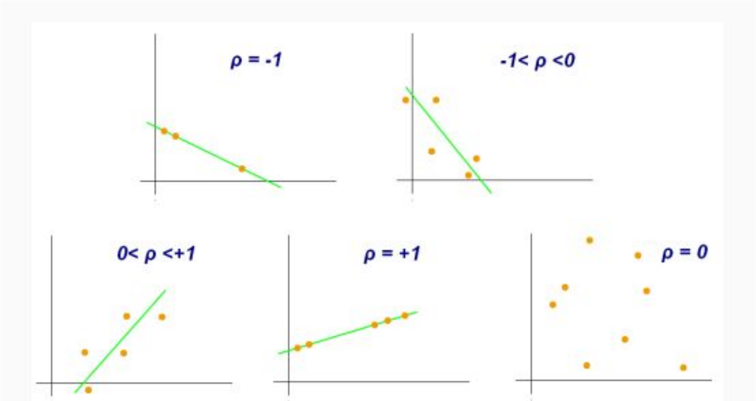
```
ggplot(ggplot2::diamonds) +
  aes(x = color) +
  geom_bar(fill = "#112446") +
  theme_minimal()
```

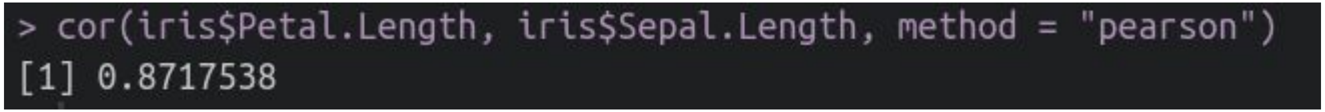
```
Listening on http://127.0.0.1:7259
```

Correlation

Pearson Correlation Coefficient (ρ):

- measures the linear correlation between two variables;
- it has a value between +1 and -1.



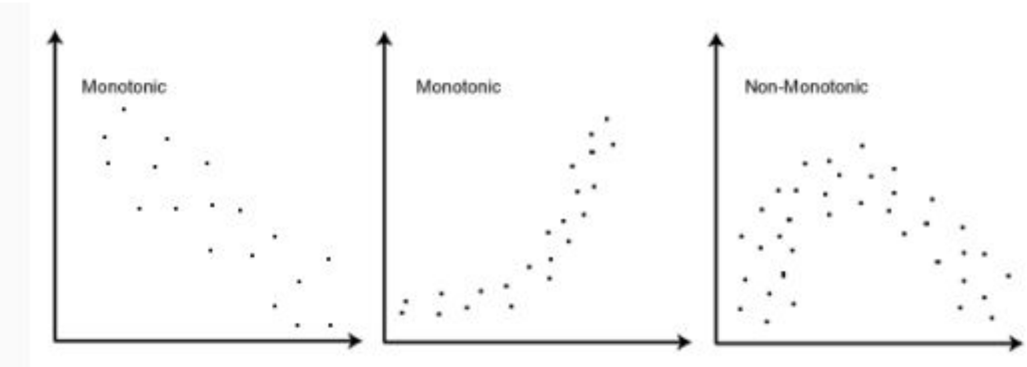


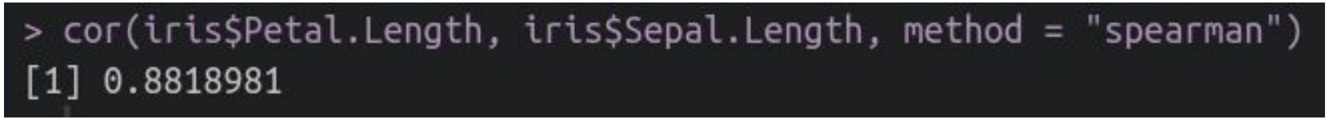
```
> cor(iris$Petal.Length, iris$Sepal.Length, method = "pearson")
[1] 0.8717538
```

Correlation

Spearman Rank-Order Correlation Coefficient:

- measures the strength and direction of monotonic association between two variables;
- two variables can be related according to a type of non-linear but still monotonic relationship.





```
> cor(iris$Petal.Length, iris$Sepal.Length, method = "spearman")
[1] 0.8818981
```

```
cor(datasets::iris[,c(1,2,3,4)])
```

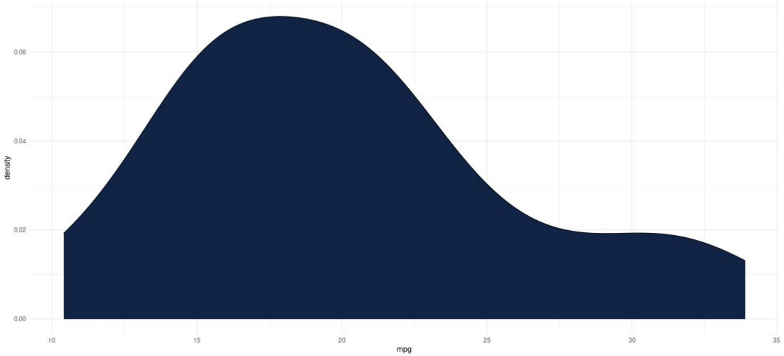
Hypothesis Tests

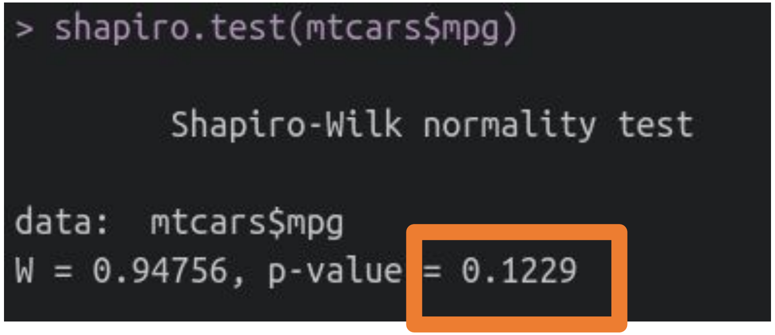
Does this sample of data follows a normal distribution?

The null and alternative hypothesis of an Shapiro-Wilk are:

H0: test is that the population is normally distributed.

H1: test is that the population isn’t normally distributed.





```
> shapiro.test(mtcars$mpg)

        Shapiro-Wilk normality test

data:  mtcars$mpg
W = 0.94756, p-value = 0.1229
```

A statistically significant test result (P > 0.05) means that the test hypothesis should not be rejected.

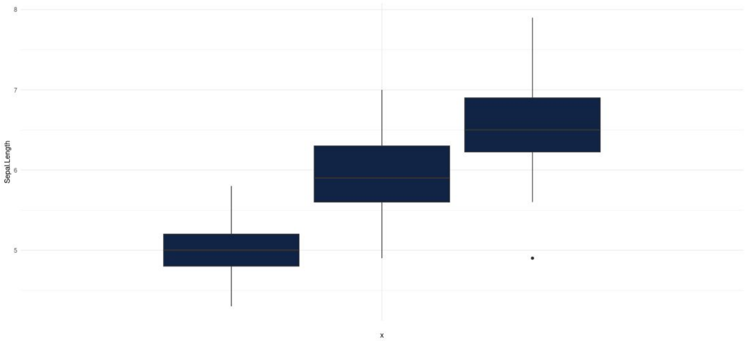
Hypothesis Tests

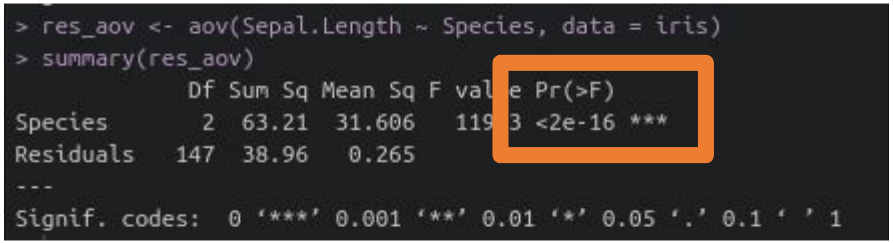
ANOVA (one-sided) to help us answer the question: “Is the length sepal 3 species of iris?”.

The null and alternative hypothesis of an ANOVA are:

H0: μ setosa=μ versicolor=μ virginica (⇒ the 3 species are equal in terms of Sepal length)

H1: *at least* one mean is different (⇒ at least one species is different from the other 2 species in terms of Sepal length)





```
> res_aov <- aov(Sepal.Length ~ Species, data = iris)
> summary(res_aov)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

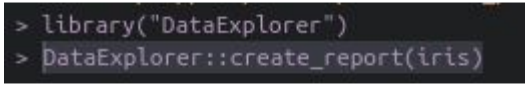
A statistically significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected. A P value greater than 0.05 means that no effect was observed.

Data Explorer

DataExplorer create reports about a data.frame

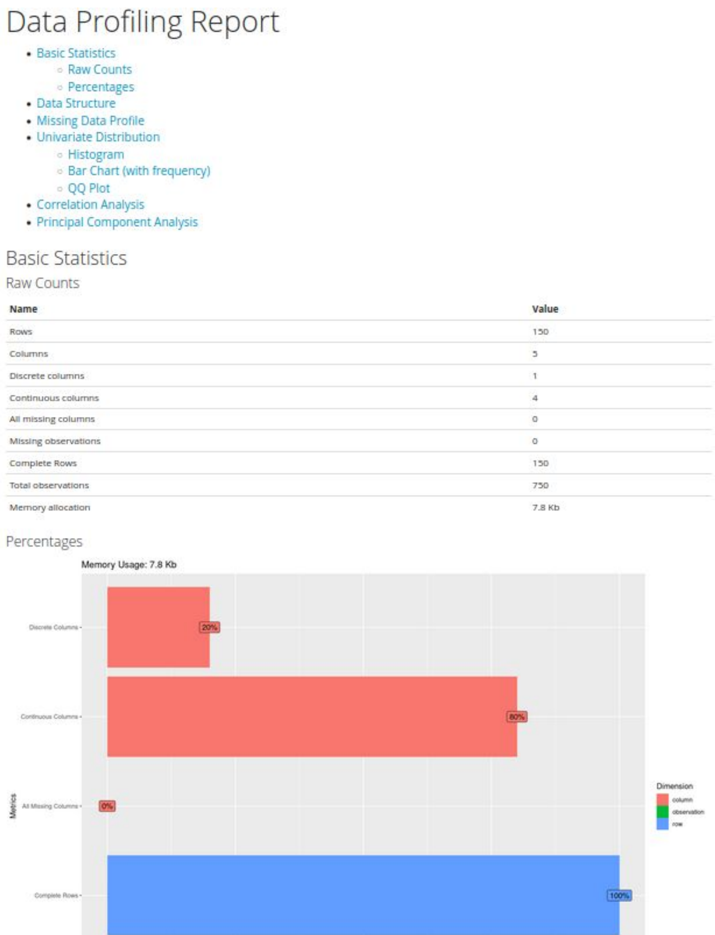
There are 3 main goals for DataExplorer:

1. **Exploratory Data Analysis (EDA)**
2. **Feature Engineering**
3. **Data Reporting**



```
> library("DataExplorer")
> DataExplorer::create_report(iris)
```

https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html



Data Profiling Report

- Basic Statistics
  - Raw Counts
  - Percentages
- Data Structure
- Missing Data Profile
- Univariate Distribution
  - Histogram
  - Bar Chart (with frequency)
  - QQ Plot
- Correlation Analysis
- Principal Component Analysis

Basic Statistics

Raw Counts

Name	Value
Rows	150
Columns	5
Discrete columns	1
Continuous columns	4
All missing columns	0
Missing observations	0
Complete Rows	150
Total observations	750
Memory allocation	7.8 Kb

Percentages