Intro to R

Ezra Garcia | ASTRAI | W07

27.11.2024

What is R?

- An open-source programming language and software environment.
- Developed by statisticians Ross Ihaka and Robert Gentleman in the early 1990s and has since become widely used in academia, research, and industry.
- Used for statistical computing, data analysis, and graphical representation.

https://osbsoftware.com.br/produto/rstudio

Common use cases

- **Data Analysis**: R allows users to manipulate, analyze, and interpret large datasets efficiently.
- **Statistical Modeling**: It provides a wide range of statistical techniques such as linear and nonlinear modeling, classification, clustering, and time-series analysis.
- **Data Visualization**: With powerful libraries like `ggplot2`, R can create high-quality plots, graphs, and charts.
- **Machine Learning**: R offers a range of machine learning algorithms for classification, regression, clustering, and dimensionality reduction.

Field-specific applications

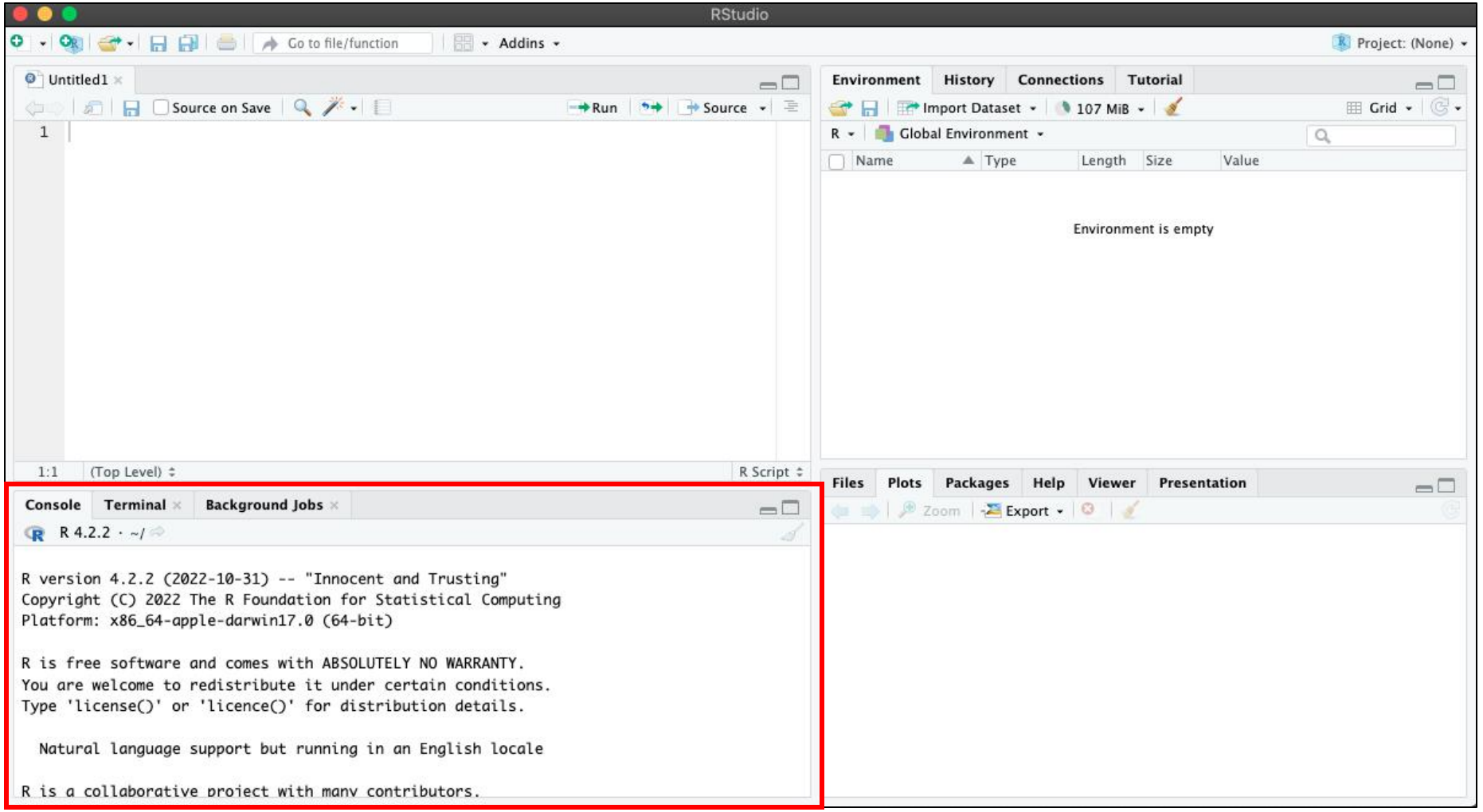
- **Bioinformatics**: R is commonly used in bioinformatics for analyzing biological data like gene expression datasets.
- **Finance**: Financial analysts use R for quantitative finance, portfolio management, and risk analysis.
- **Econometrics**: Economists use R for analyzing economic data, building econometric models, and conducting hypothesis testing.
- **Transportation**: R can be leveraged for traffic flow analysis, route optimization, demand forecasting, crash data analysis, spatial analysis, survey analysis, etc.

What makes R special?

- **Open Source**: R is free to use.
- **Comprehensive Libraries**: R offers a multitude of packages (e.g., `ggplot2`, `dplyr`, `caret`, `shiny`) for various types of data analysis and visualization.
- **Cross-Platform**: R works across various platforms (Windows, Mac, Linux).
- **Extensive Community Support**: R has an active user community, meaning constant improvements and a wealth of resources.
- **High-Quality Graphics**: R's visualization capabilities allow for the creation of publication-ready graphics.
- **Flexibility with Data**: R can handle various types of data, including structured (CSV, Excel) and unstructured data (text, images).

The RStudio user-interface

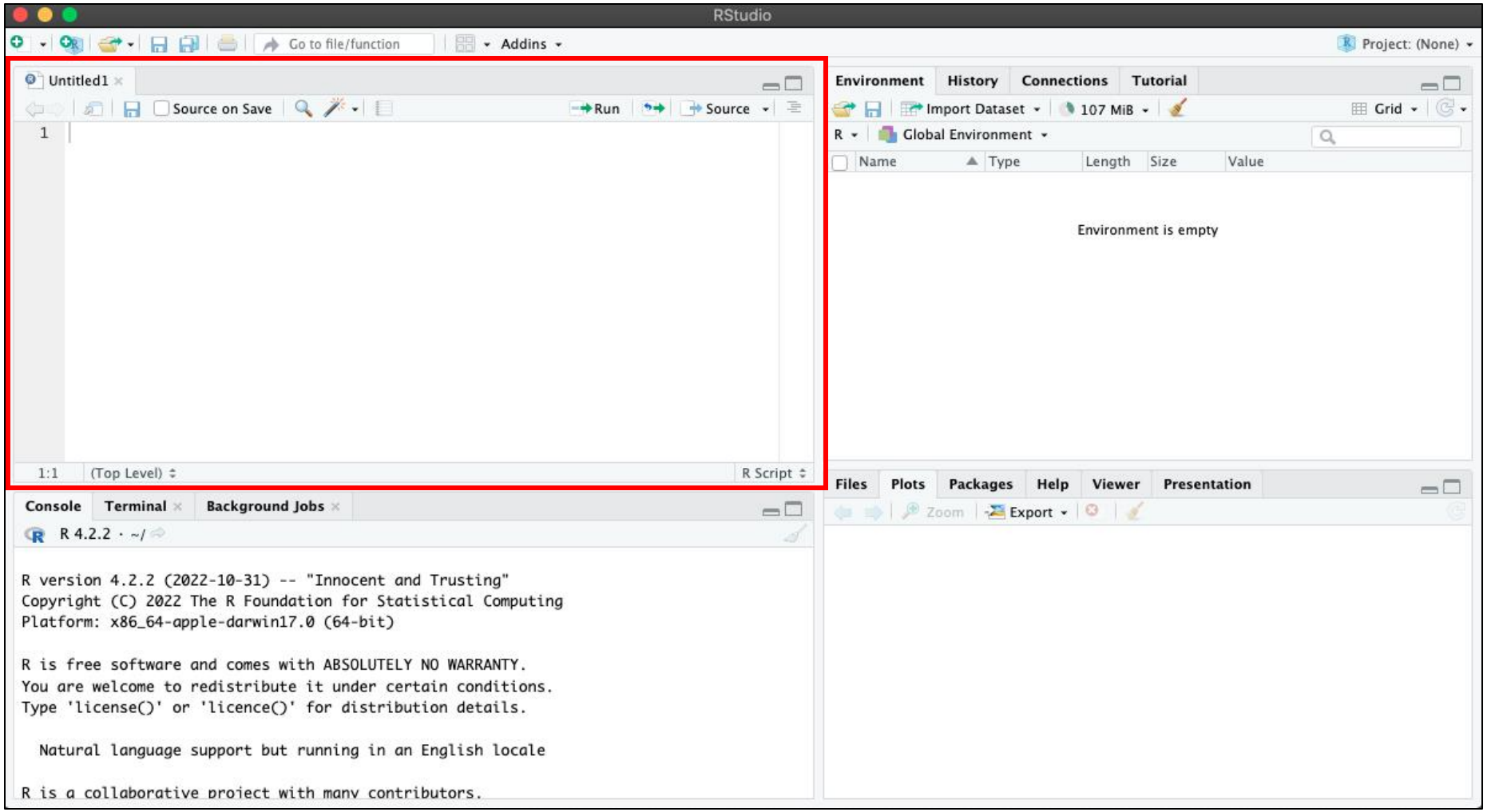
Console window: where output and computations are displayed.



Source: Screenshot taken by E. Garcia

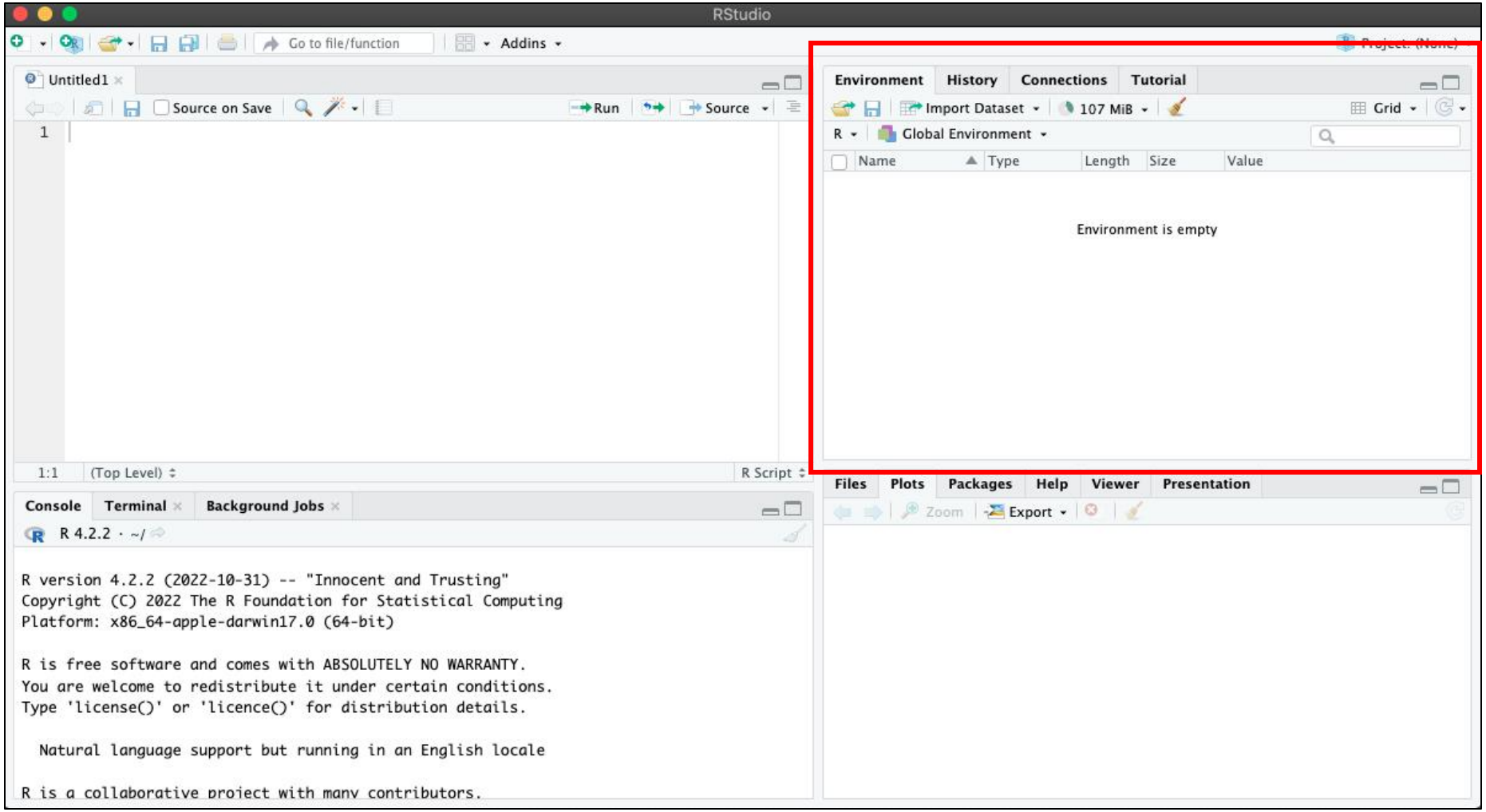
The RStudio user-interface

Source (code-editing) window: used for editing a script.



Source: Screenshot taken by E. Garcia

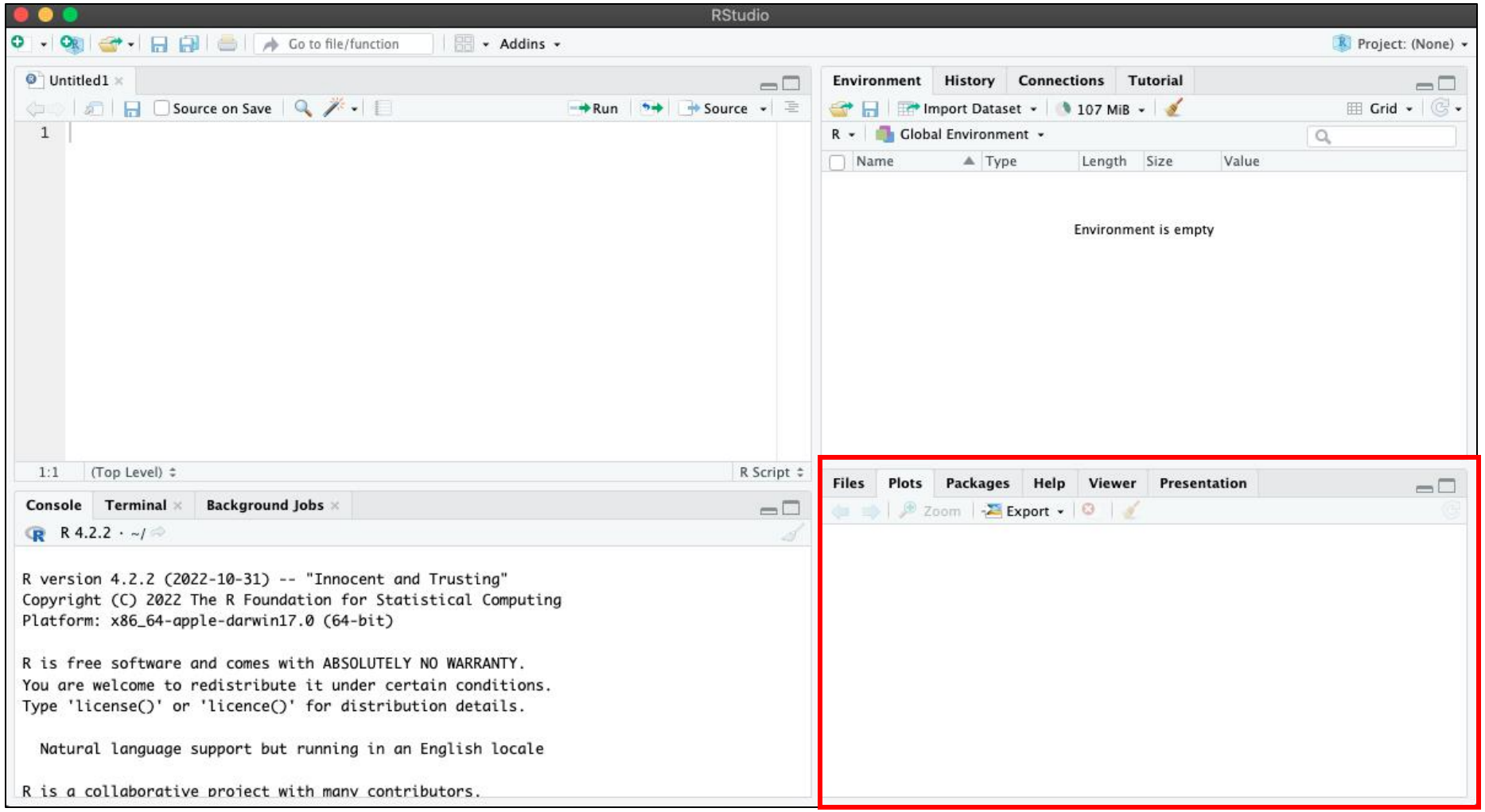
The RStudio user-interface



Workspace and History window: where objects available for computations are displayed.

Source: Screenshot taken by E. Garcia

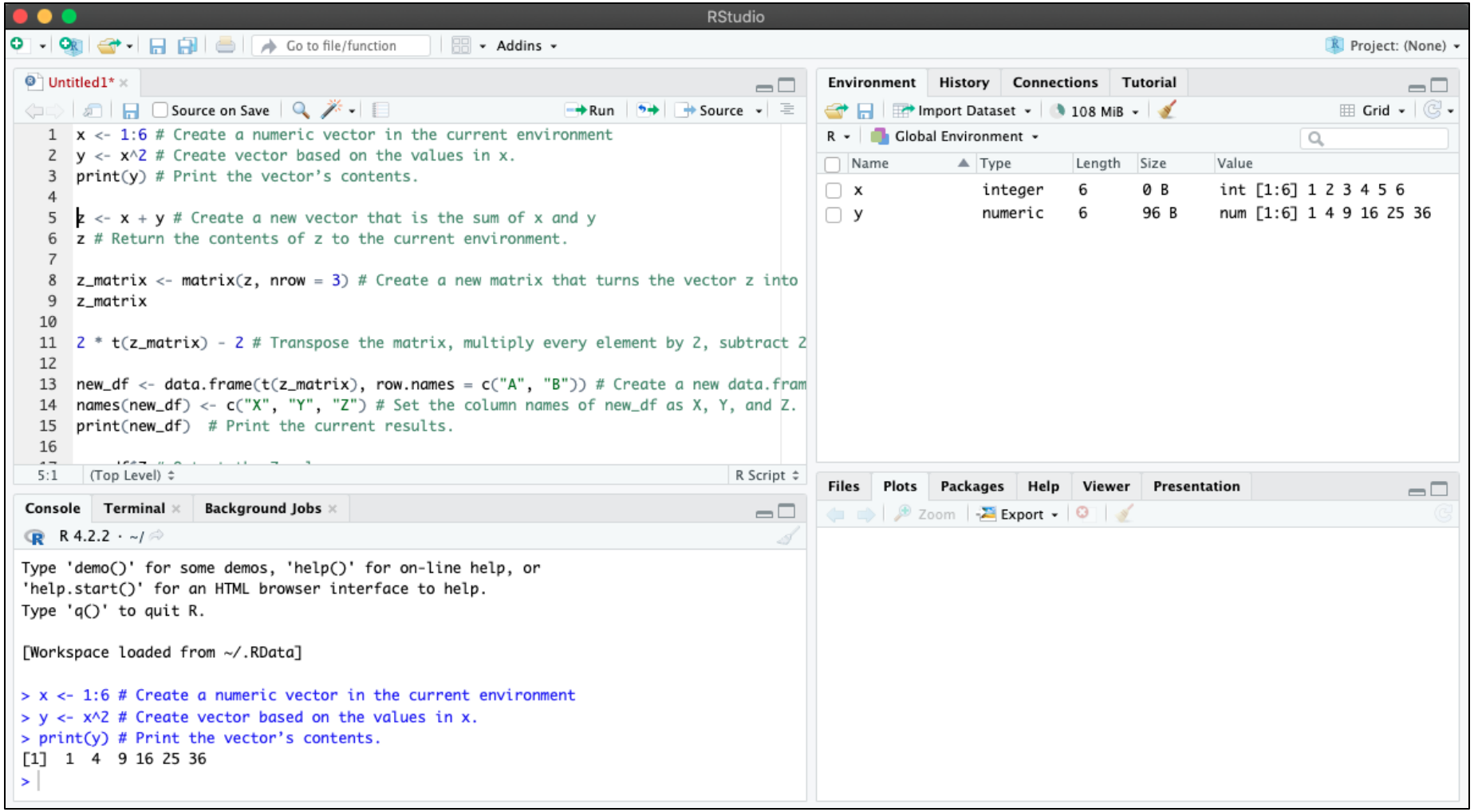
The RStudio user-interface



Source: Screenshot taken by E. Garcia

The Plots and Files window: multi-use panes.

The RStudio user-interface



Source: Screenshot taken by E. Garcia

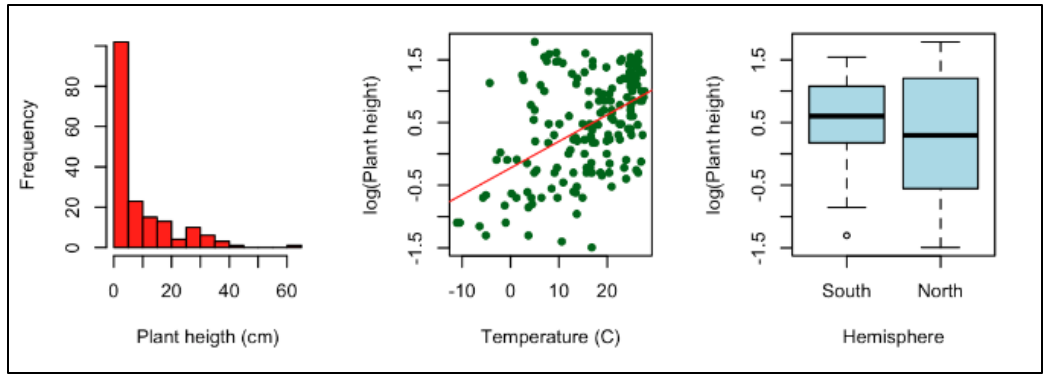
Basic commands

Assigning values	Creating vectors	Matrices
• x <- 5 • y = 10	• v <- c(1, 2, 3, 4) • v[2] • v[v > 3]	• m <- matrix(1:9, nrow=3) • m[1,] • m[, 2]

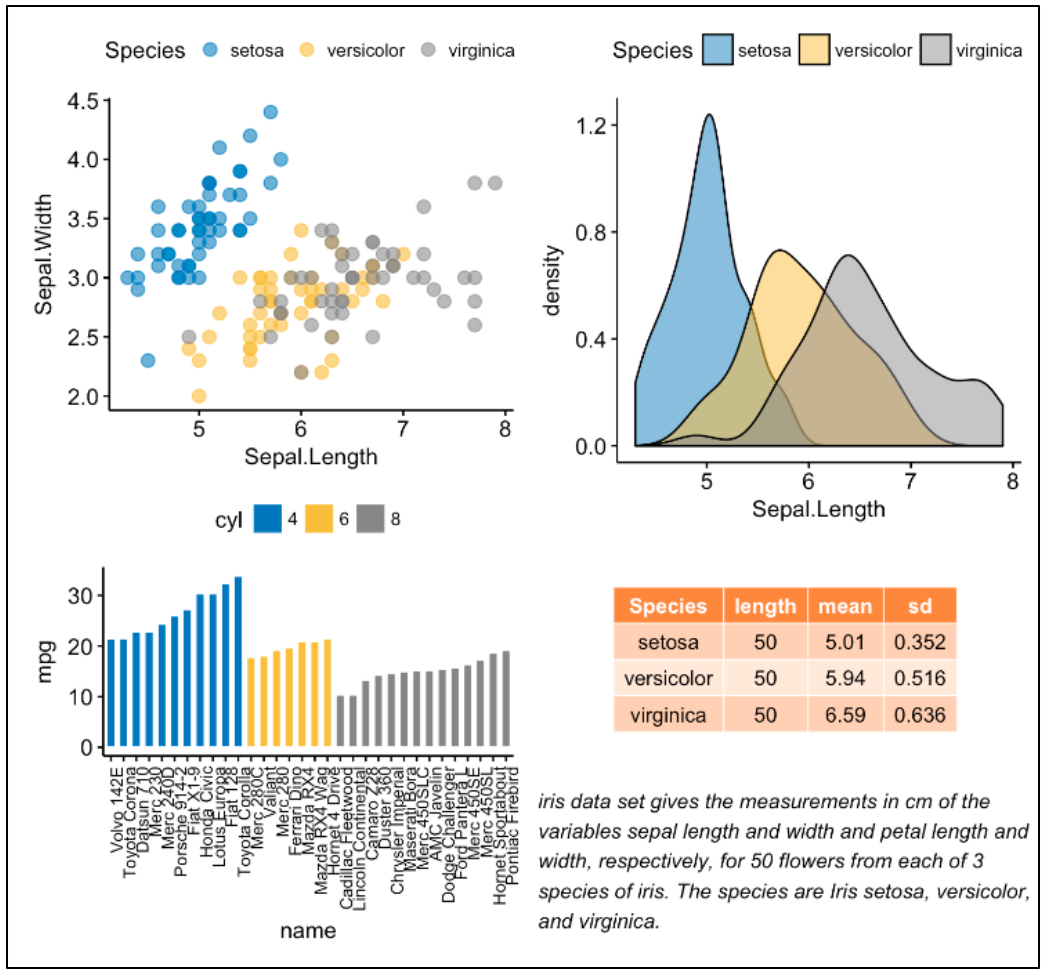
Basic statistical functions	Data Frames	Installing/loading packages
• mean(v) • median(v) • sd(v) • summary(v)	• df <- data.frame(Name=c('Alice', 'Bob'), Age=c(25, 30)) • df$Name • df[1,]	• install.packages("ggplot2") • library(ggplot2)

Commands for visualization

- plot(v) # Plot a vector
- hist(v) # Histogram
- boxplot(v) # Boxplot



https://environmentalcomputing.net/graphics/basic-plotting/



Species	length	mean	sd
setosa	50	5.01	0.352
versicolor	50	5.94	0.516
virginica	50	6.59	0.636

iris data set gives the measurements in cm of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

https://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/

Packages

Data Manipulation	Data Visualization	Statistical Analysis
• **`dplyr`**: For data manipulation (e.g., filtering, summarizing). • **`tidyr`**: For reshaping and tidying data. • **`data.table`**: For fast and efficient manipulation of large datasets.	• **`ggplot2`**: For creating sophisticated visualizations. • **`plotly`**: For interactive plots. • **`lattice`**: For multivariate data visualization.	•**`stats`**: Built-in package for basic statistical methods. •**`car`**: For regression analysis and diagnostics. •**`psych`**: For psychological and multivariate statistics.
Machine Learning	**Time Series Analysis**	**Other**
• **`caret`**: Comprehensive suite for machine learning workflows. • **`randomForest`**: For random forest models. • **`xgboost`**: For gradient boosting models.	• **`forecast`**: For forecasting models like ARIMA. • **`zoo`**: For managing and analyzing time series data.	•**`shiny`**: For building interactive web applications. •**`tm`**: For text mining and natural language processing. •**`sf`**: For spatial data handling.

How can R be used in transportation research? Strengths that other languages don't offer?

- R excels in statistical computing, making it ideal for analyzing transportation-related data where statistical modeling and hypothesis testing are required.
- R provides high-quality, publication-ready graphics. Packages like `ggplot2`, `shiny`, and `plotly` allow detailed visualizations of transportation networks and traffic patterns.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

https://www.statsdirect.co.uk/help/parametric_methods/utt.htm

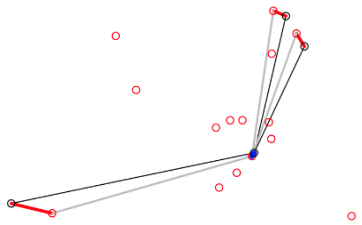
```
> t.test(data$X, data$Y)   # T-test to compare means of X and Y

        Welch Two Sample t-test

data:  data$X and data$Y
t = -22.966, df = 158.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -52.04668 -43.80368
sample estimates:
mean of x mean of y
 50.32515  98.25033
```

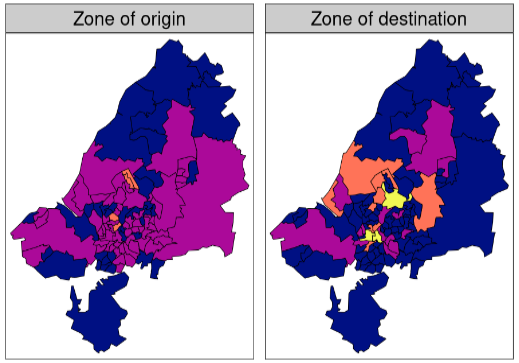
Source: Screenshot taken by E. Garcia

Feature	R	Python
Statistical Modeling	R is unparalleled in statistical methods, making it the go-to for rigorous analysis.	Python relies on libraries like statsmodels and scipy, which may not be as robust.
Geospatial Analysis	Specialized tools like sf and tmap offer deeper support for geospatial data.	Libraries like GeoPandas and folium are powerful but less mature than R's.
Data Visualization	ggplot2 and shiny allow for high-quality, interactive visualizations.	Python's matplotlib and seaborn are strong, but less intuitive for beginners.
Ease of Use for Statistics	R is designed with statistical applications in mind, making it more intuitive.	Python requires more effort to set up equivalent workflows for statistics.
General-Purpose Programming	Less versatile for general programming outside data science.	A general-purpose language, Python is better for tasks beyond analysis.
Community	Strong in academia and research (especially social sciences and transportation).	Dominant in industry, with extensive support for machine learning and automation.

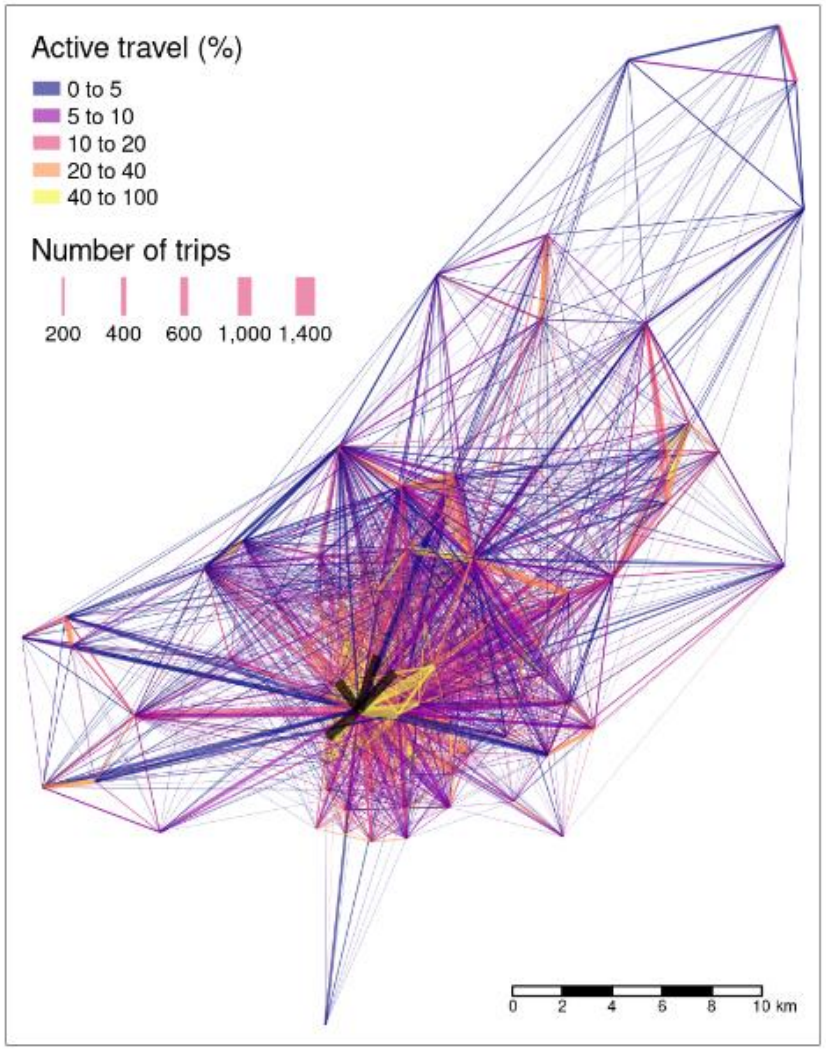
A transportation case study

A Bristol, UK case study aim at reducing motorized vehicle congestion and encouraging bicycle traffic.





o	d	all	bicycle	foot	car_driver	train
E02003043	E02003043	1493	66	1296	64	8
E02003047	E02003043	1300	287	751	148	8
E02003031	E02003043	1221	305	600	176	7
E02003037	E02003043	1186	88	908	110	3
E02003034	E02003043	1177	281	711	100	7



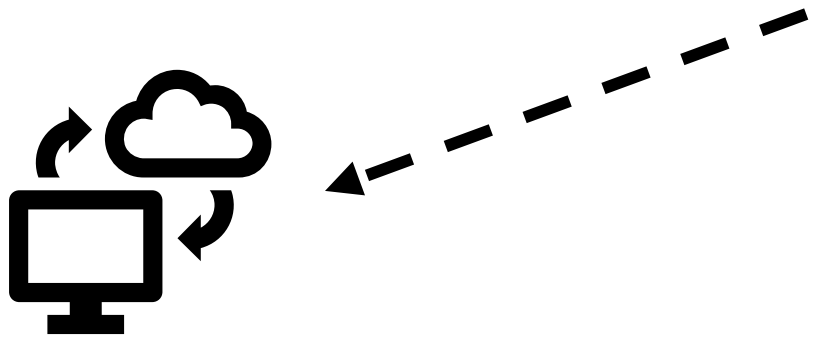
https://bookdown.org/robinlovelace/geocompr/transport.html

Links to installation

- R: https://cran.r-project.org/
- RStudio: https://posit.co/download/rstudio-desktop/

Learning by doing

- Download R file "syntax.R" from GitHub
- Download R file "practice.R" from GitHub
- Open files in RStudio
- Run scripts



README

ASTRAI

The Art and Science of Transportation Research in the AI Era

6-CP seminar (Winter Semester 2024/25)
Wednesdays 11:40-15:00
L402/231

References

- https://cran.r-project.org/doc/manuals/R-intro.pdf
- https://developer.r-project.org/equalAssign.html
- https://bookdown.org/robinlovelace/geocompr/transport.html