QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 – 2019/2020

This is R

QUEEN'S UNIVERSITY BELFAST | SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

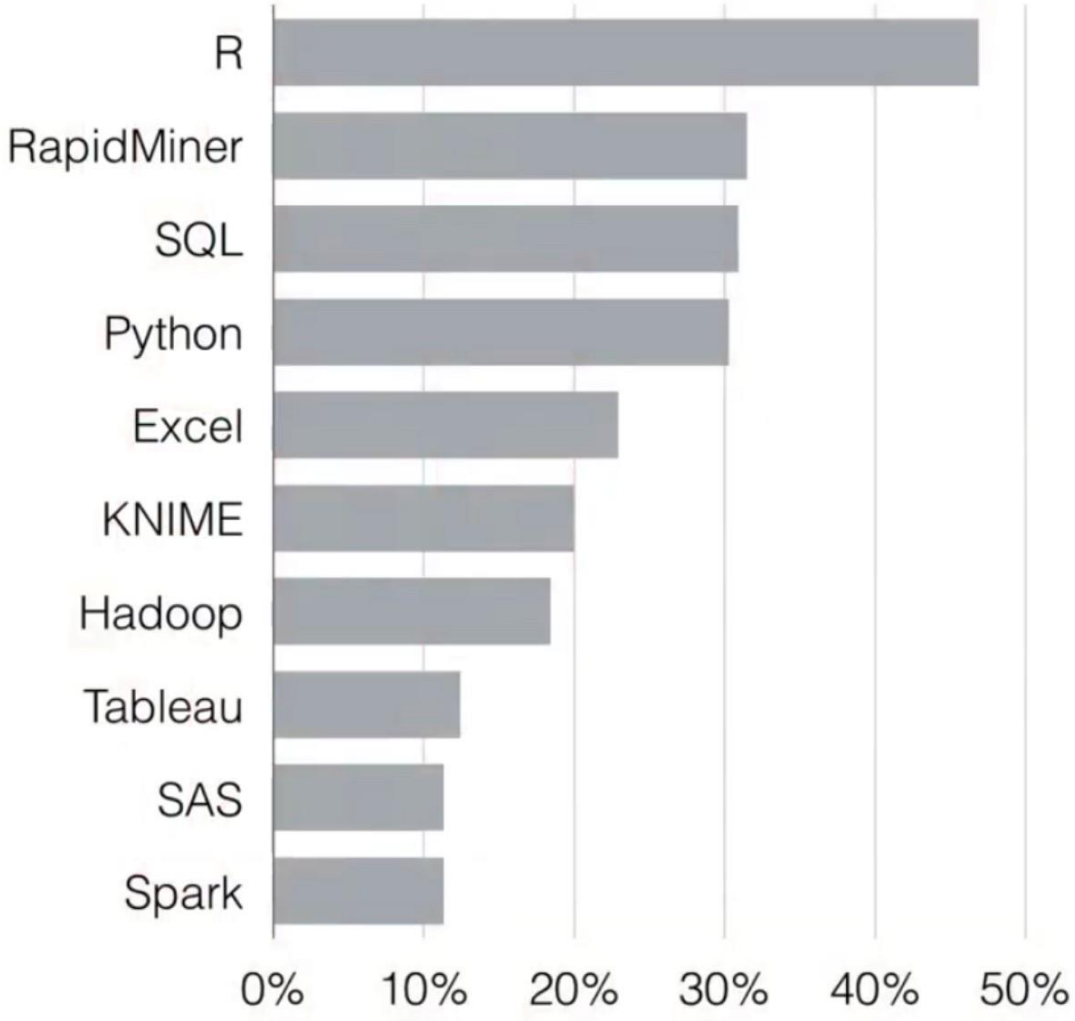
This is R

```
###################################################################################################
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
###################################################################################################
# Start
#"""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""
# Clustering 450K methylation CpGs probes written by Dr Reza Rafiee
# Research Associate, Northern Institute for Cancer Research, Newcastle University
# This script loads 20,000 methylation probes (from 450K methylation profiling) and doing clustering analysis
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#

library(mclust) # Gaussian Mixture Modelling package for Model-Based Clustering, Classification, and Density Estimation
library(scatterplot3d)
library(pheatmap)
library(apcluster) # Affinity Propagation Clustering

load("~/20KBetaValues_51InfantSHH.RData")     # 20,000 probes
length(colnames(BetaValues_51Samples_20K))    # n=51

# Performs a principal components analysis on the given data matrix and returns the results as an object of class prcomp
PCA_Comp_Scaled_Centered <- prcomp(t(BetaValues_51Samples_20K), center = TRUE, scale=T)  # scale =T is appropriate for high-dimens
summary(PCA_Comp_Scaled_Centered)
```

This is R

The language of data science

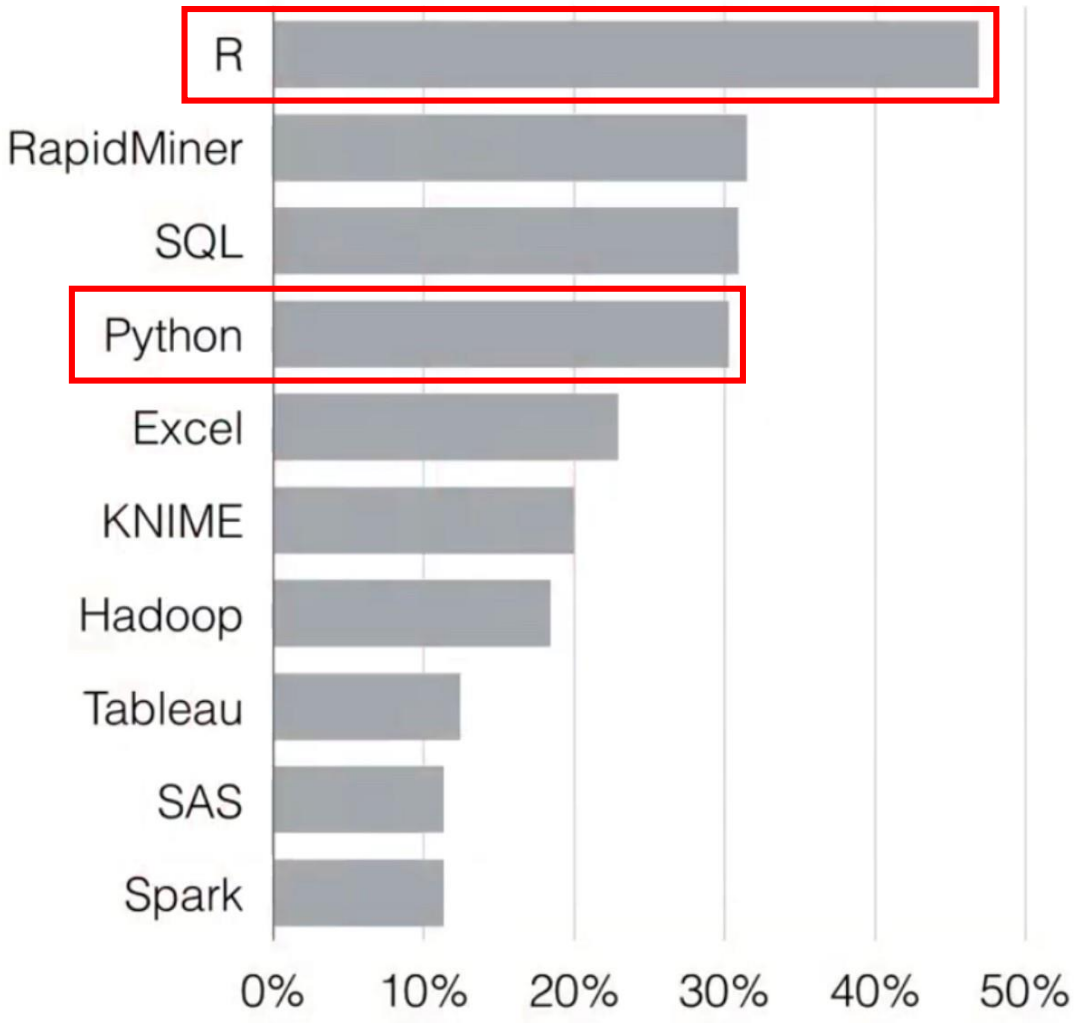
R ranking



Ranking

- Survey of data mining experts
- R is first
- 50% more use than Python

R ranking



Ranking

- Survey of data mining experts
- R is first
- 50% more use than Python

R

Free & open source

Vector operations

Great community

9000+ packages

R Programming – general properties

- An **open source** programming language freely available under the GNU public license (written in primarily C and Fortran). It's an **interpreted** language.
- A **powerful tool** to work with data (**statistical computing** and **data analysis**).
- **Highly extensible** through the use of user-submitted **packages** for specific functions or specific areas of study.
- Including **thousands of packages**, designed, maintained, and widely used by **data scientist** and **statisticians**.
- Advanced users can write C, C++, Java, .NET or Python code to manipulate R objects directly.
- **Portable** and works equally well on Windows, Linux, Mac OS.
- R has stronger **object-oriented programming facilities** than most statistical computing languages.
- Writing and running R scripts could be either in
  - **The command line interface** or a Graphical User Interface called **RStudio**.

R Programming; R packages

- A great advantage to the open source nature of R is that users have contributed a huge number of packages for solving a vast majority of data analysis problems.
- For example, there are packages specifically directed to **visualise data**, **non-parametric statistics**, **signal processing**, **bioinformatics** and so on.

- **Two steps** to use any packages.
  - First, they **must be installed on your system**, and this is an **one-time** step.
  - Second, once a package is installed, it **must be loaded** by calling the ***library()*** function with the name of the library as an argument.

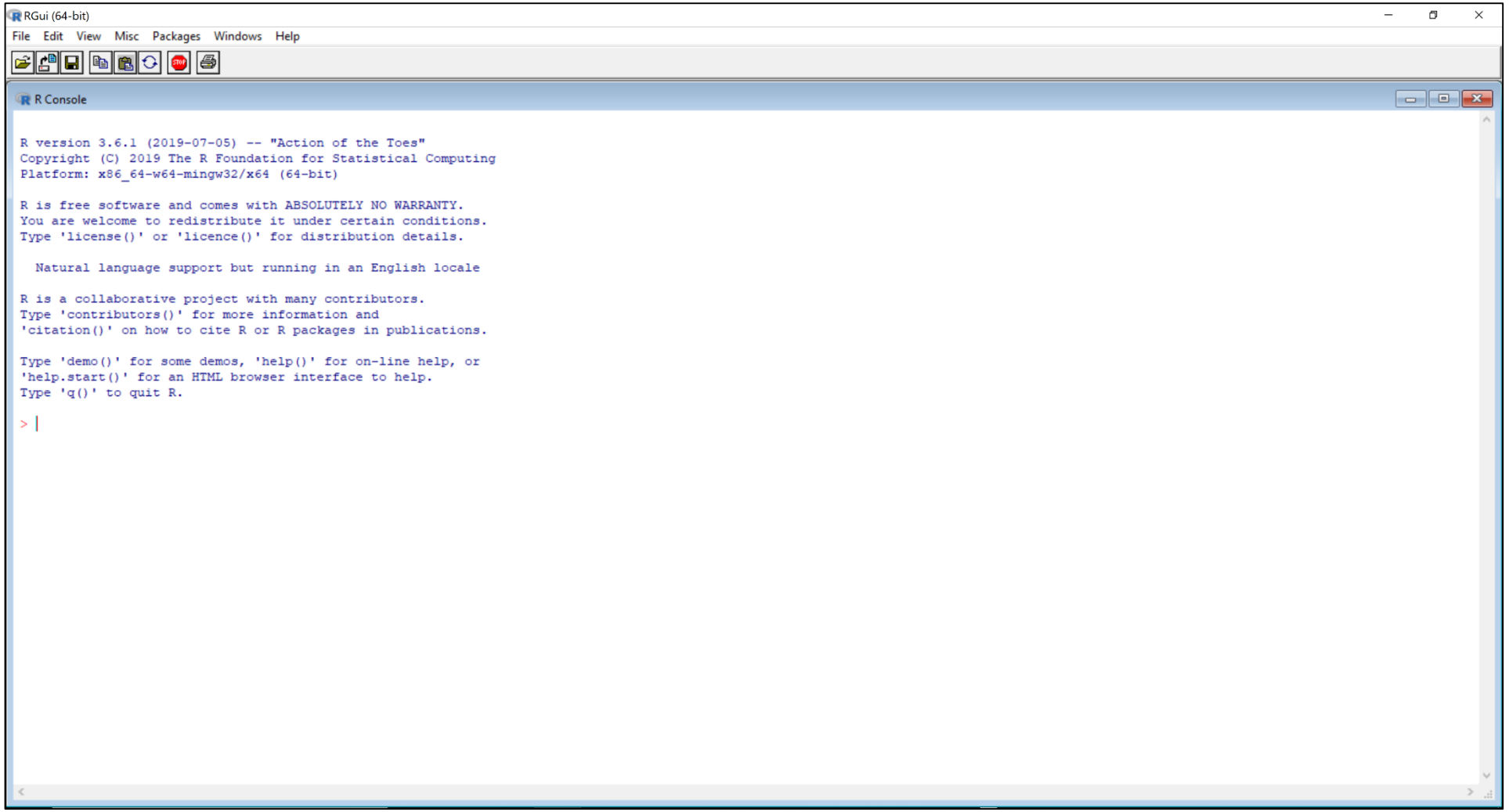
- Some examples of using installed packages in R:
  - library(NMF) # Nonnegative Matrix Factorization (Algorithms and Framework) package
  - library(mclust) # Gaussian Mixture Modelling package for Model-Based Clustering, Classification, and Density Estimation
  - library(shiny) # Web Application Framework for R

- **CRAN** (The **C**omprehensive **R** **A**rchive **N**etwork)
  - https://cran.r-project.org/

R Programming; The command line interface

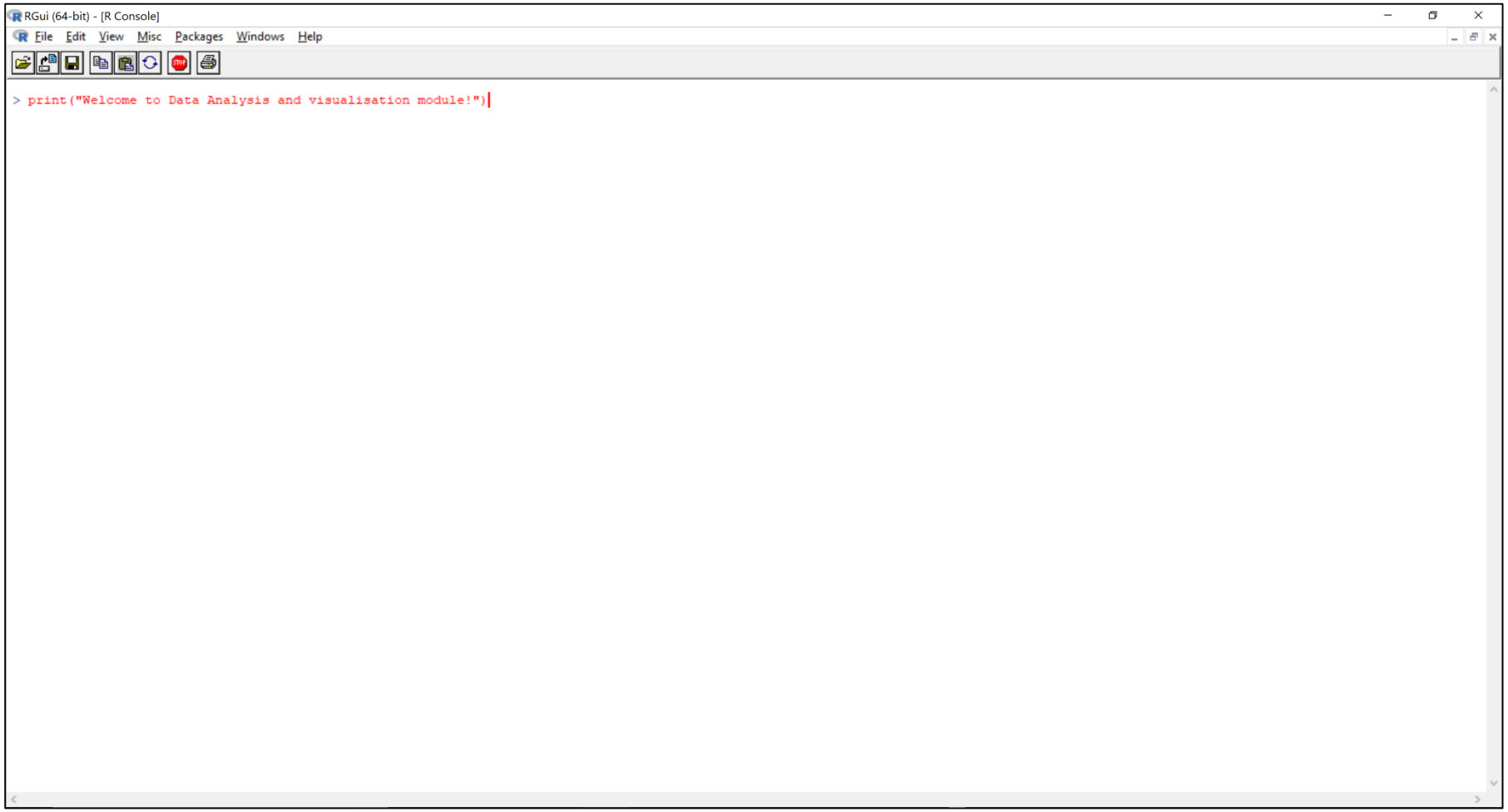
- **The command line interface**

R Console



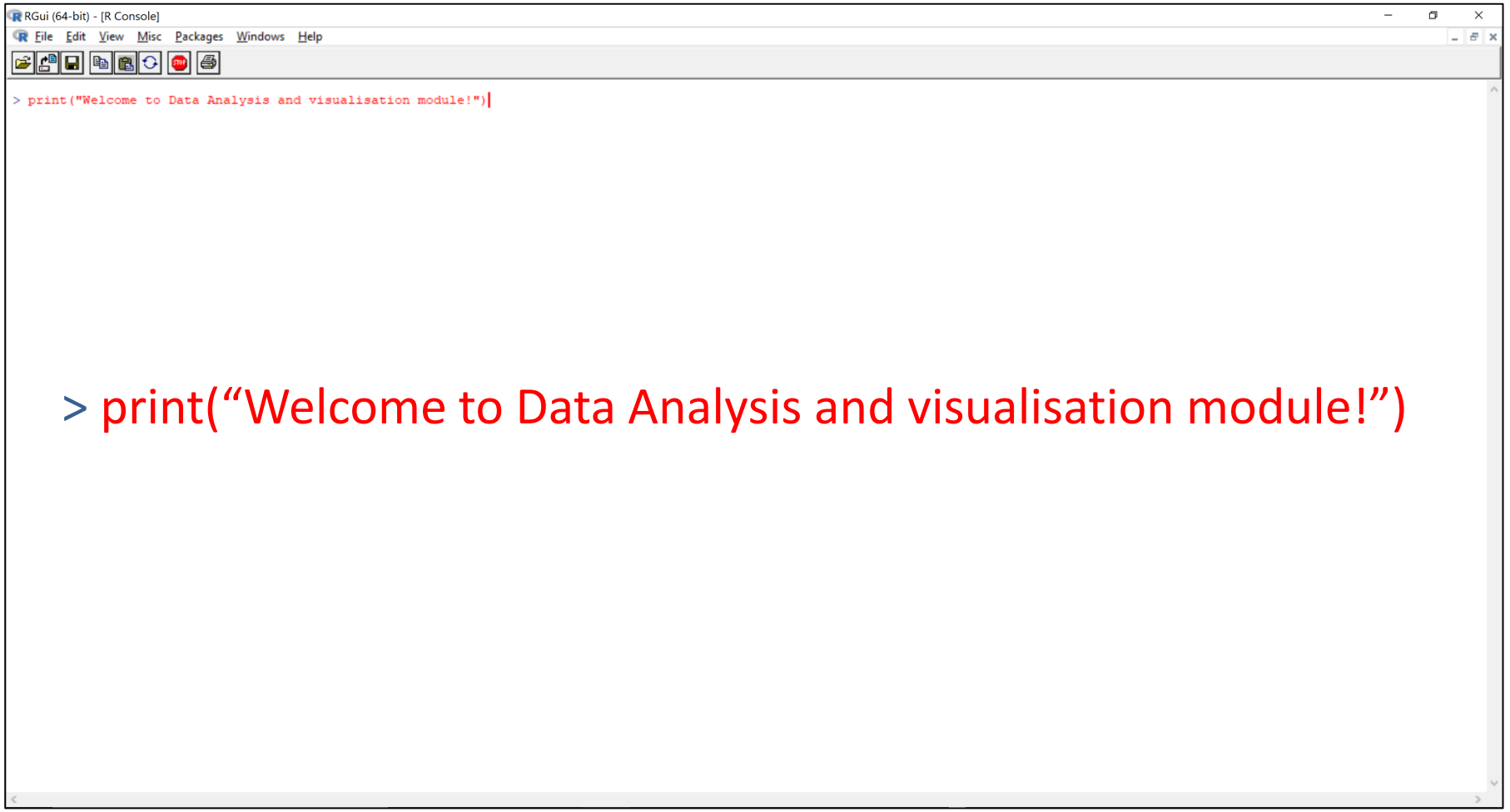
R Programming; The command line interface

- The command line interface
- To clear the console: **CTRL + Enter**



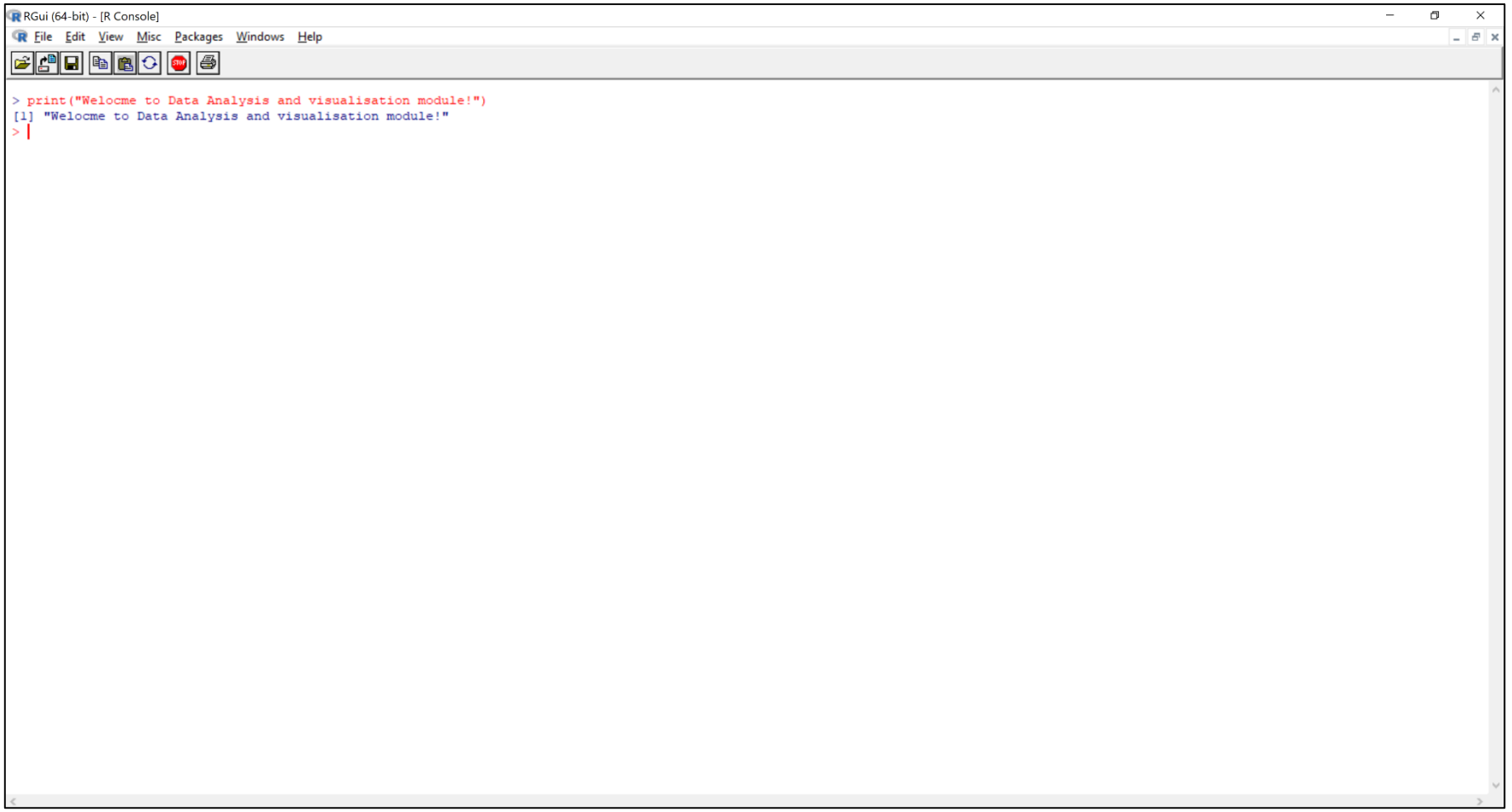
R Programming; The command line interface

- The command line interface
- To clear the console: **CTRL + Enter**



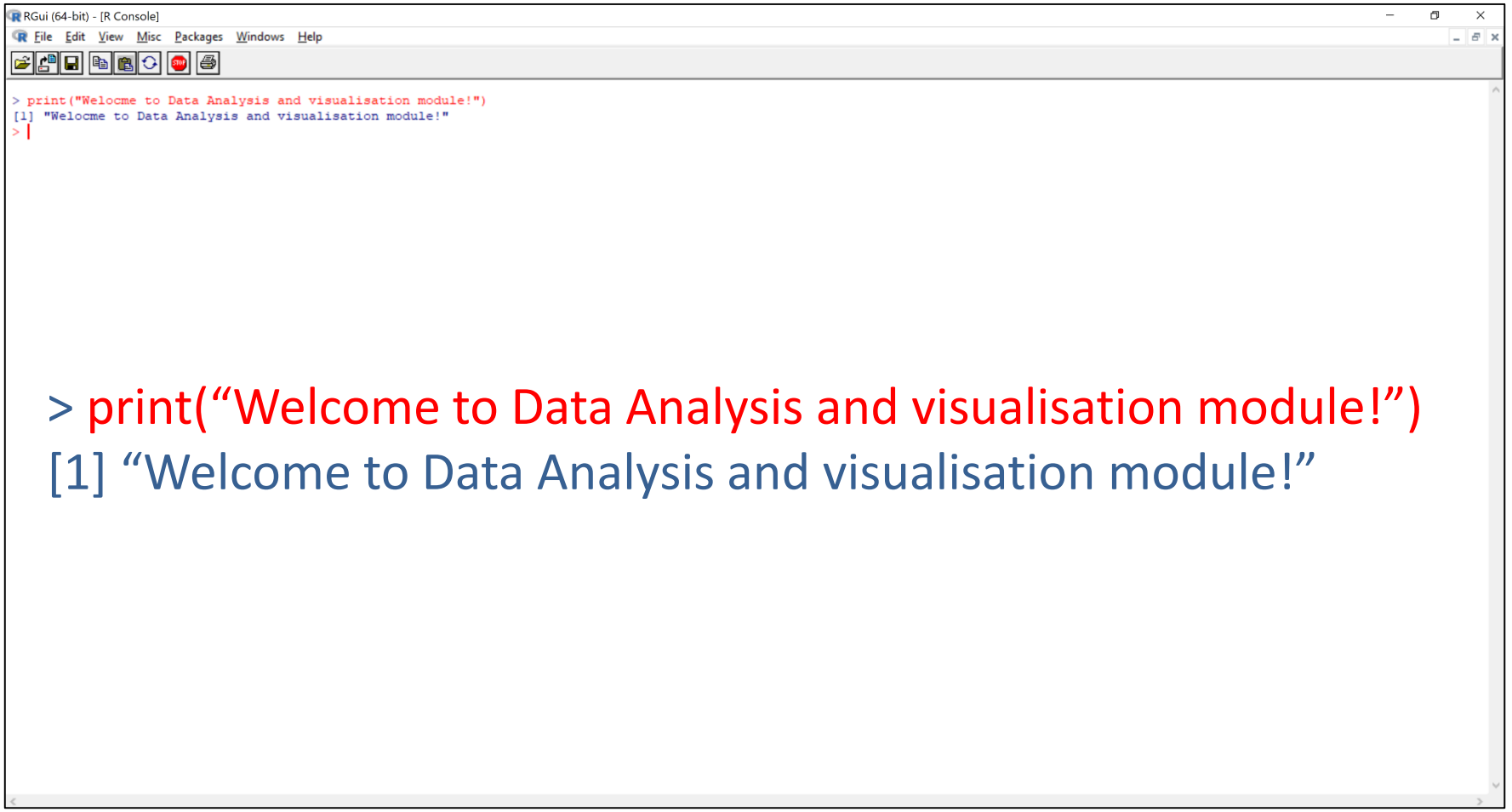
R Programming; The command line interface

- **The command line interface**
- To clear the console: **CTRL + Enter**



R Programming; The command line interface

- **The command line interface**
- To clear the console: **CTRL + Enter**



R Programming; The command line interface

- The command line interface
- R is a case-sensitive language (similar to Unix/Linux)

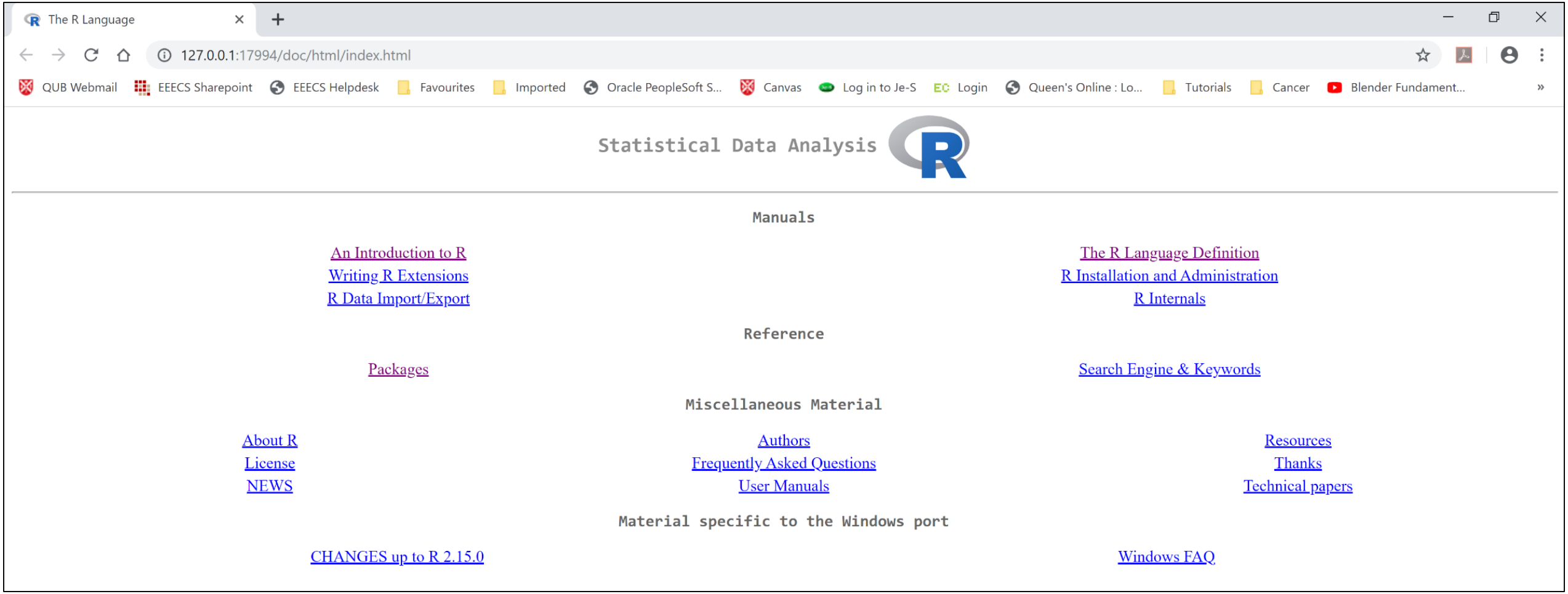
```
> Print("Welcome to Data Analysis and visualisation module!")
Error in Print("Welcome to Data Analysis and visualisation module!") :
  could not find function "Print"
>
```

R Programming; help.start() command

- **The command line interface**
- R is a case-sensitive language (similar to Unix/Linux)
- print() is an R function

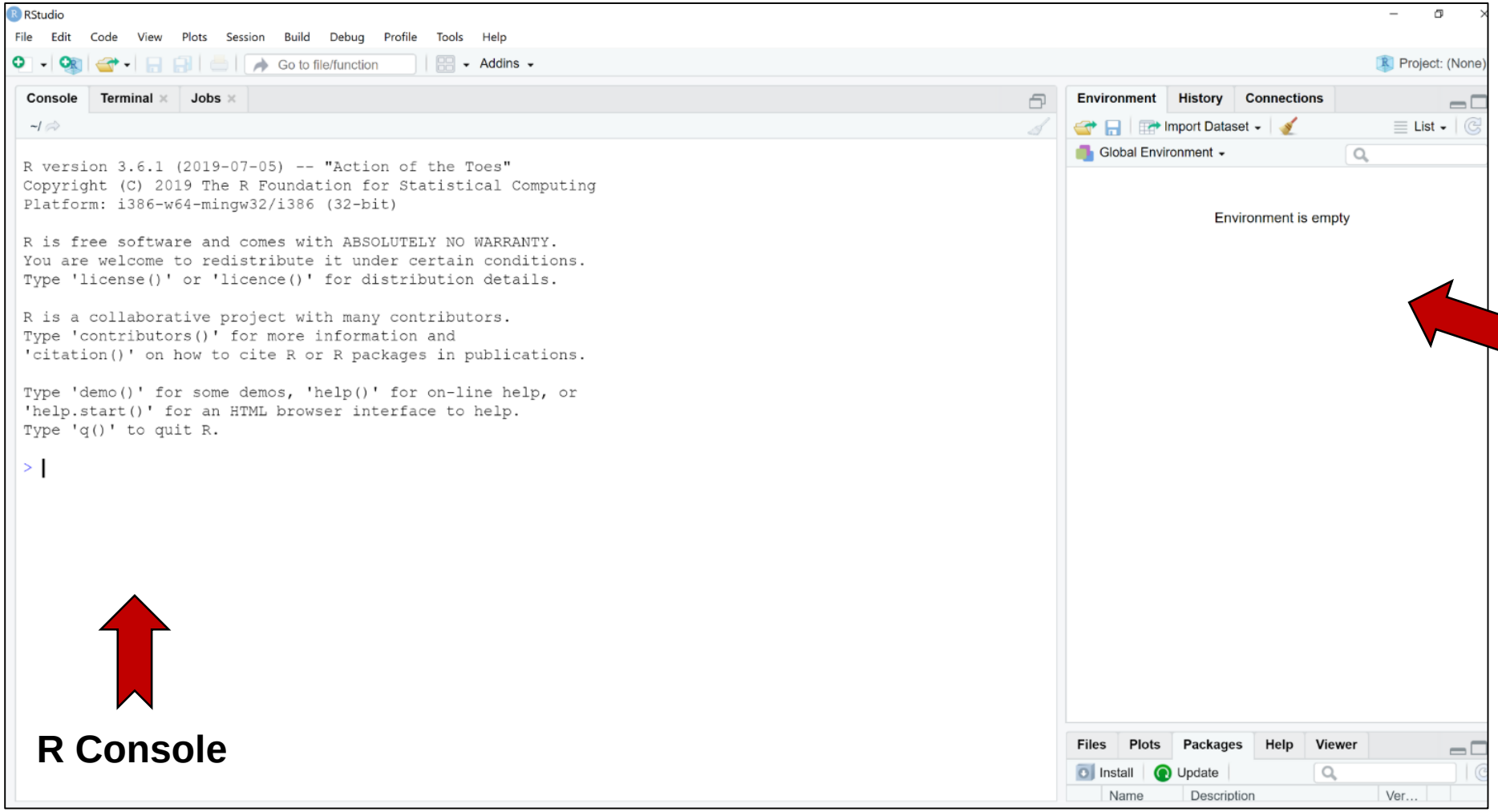
R Programming; help.start() command

- **The command line interface**
- R is a case-sensitive language (similar to Unix/Linux)
- print() is an R function
- help.start()



R Programming; RStudio

- **RStudio**



How to install R & RStudio

- R refers to a software environment that comes with a GUI (Graphical User Interface). R GUI looks more similar to the old DOS console than to SPSS or Stata.
  - For different OS (Linux, Win or Mac, etc. users): https://cran.r-project.org/
- RStudio is an IDE (Integrated Development Environment) that makes R easier to use. It includes a code editor, debugging and visualization tools.
  - For different platforms : https://rstudio.com/products/rstudio/download/
- To download and install R and RStudio (Win or Mac)
  - Step by step: ttps://courses.edx.org/courses/UTAustinX/UT.7.01x/3T2014/56c5437b88fa43cf828bff5371c6a924/

R Programming; R-object

- In contrast to other programming languages like C and java, the variables in R are **not declared** as some data type. The **variables** are assigned with **R-Objects** and the data type of the R-object becomes the data type of the variable. There are many types of R-objects.

- The frequently used **R-Objects**:
  - Vectors
  - Lists
  - Matrices
  - Arrays
  - Factors
  - Data Frames