R, RStudio, Rmarkdown

Introduction

R is a programming language

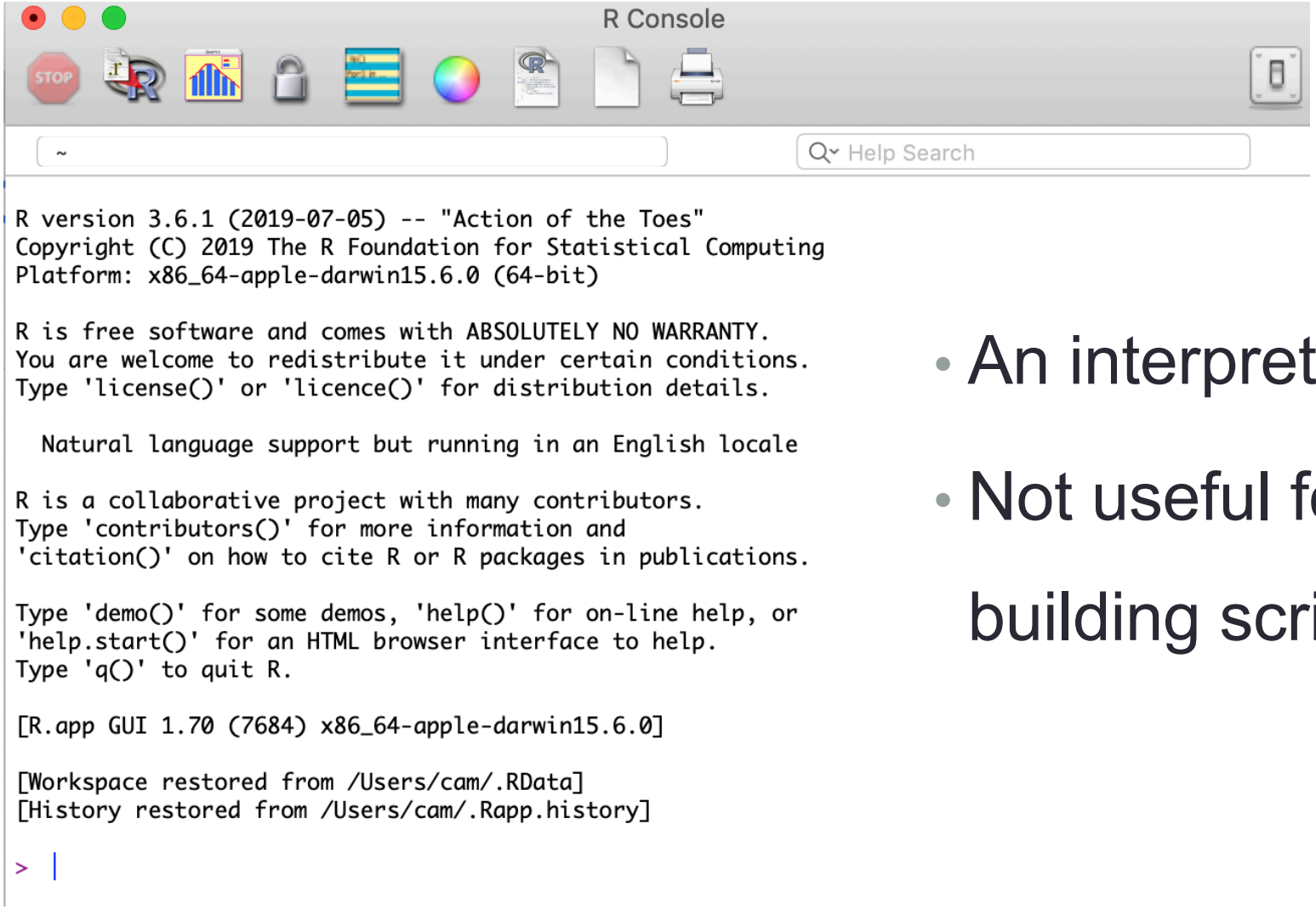
for statistical computing and graphical display

Introduction

When installing R we get

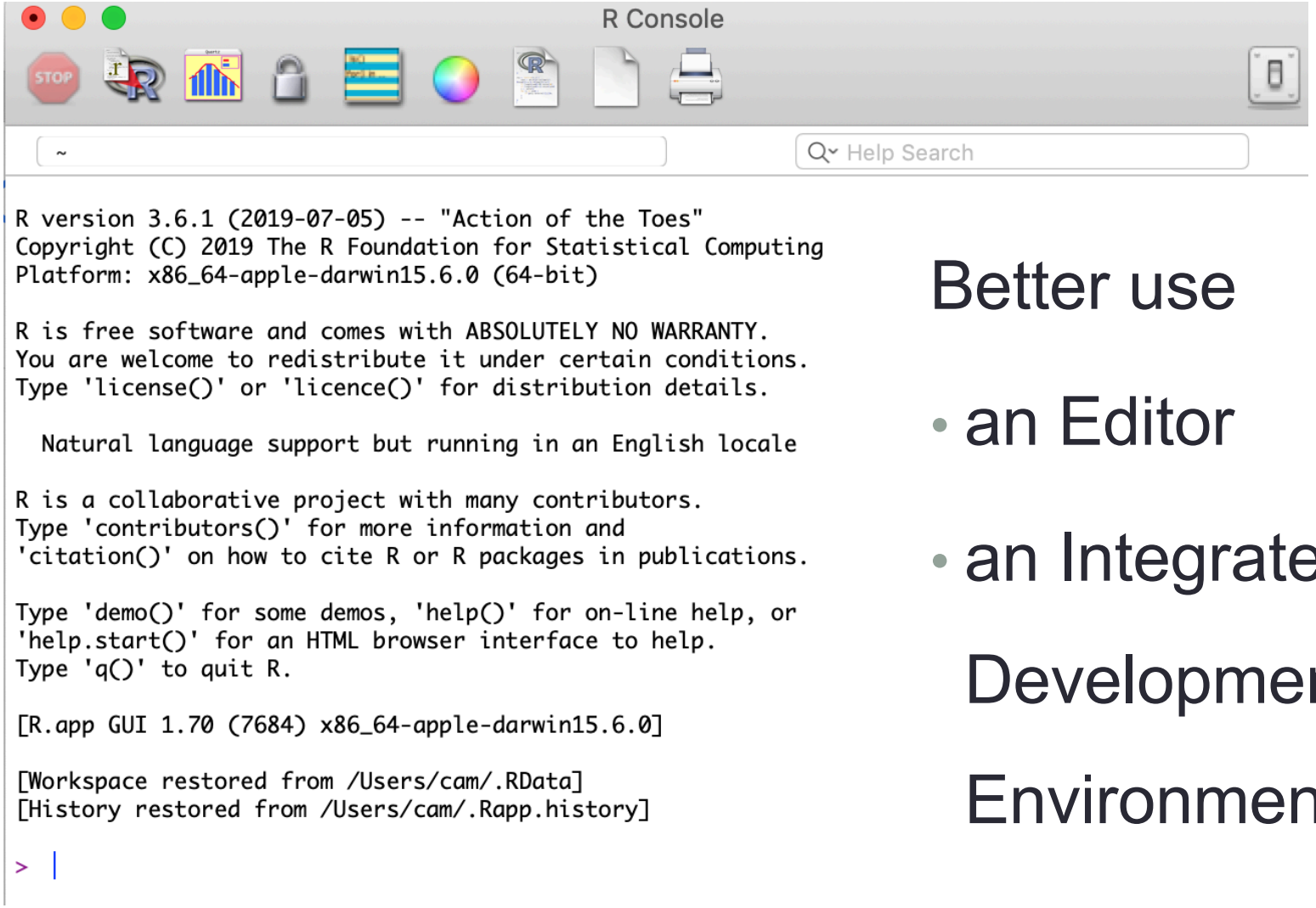
- R console
- R interpreter

R console



- An interpreter for **R**
- Not useful for building scripts

R console



Better use

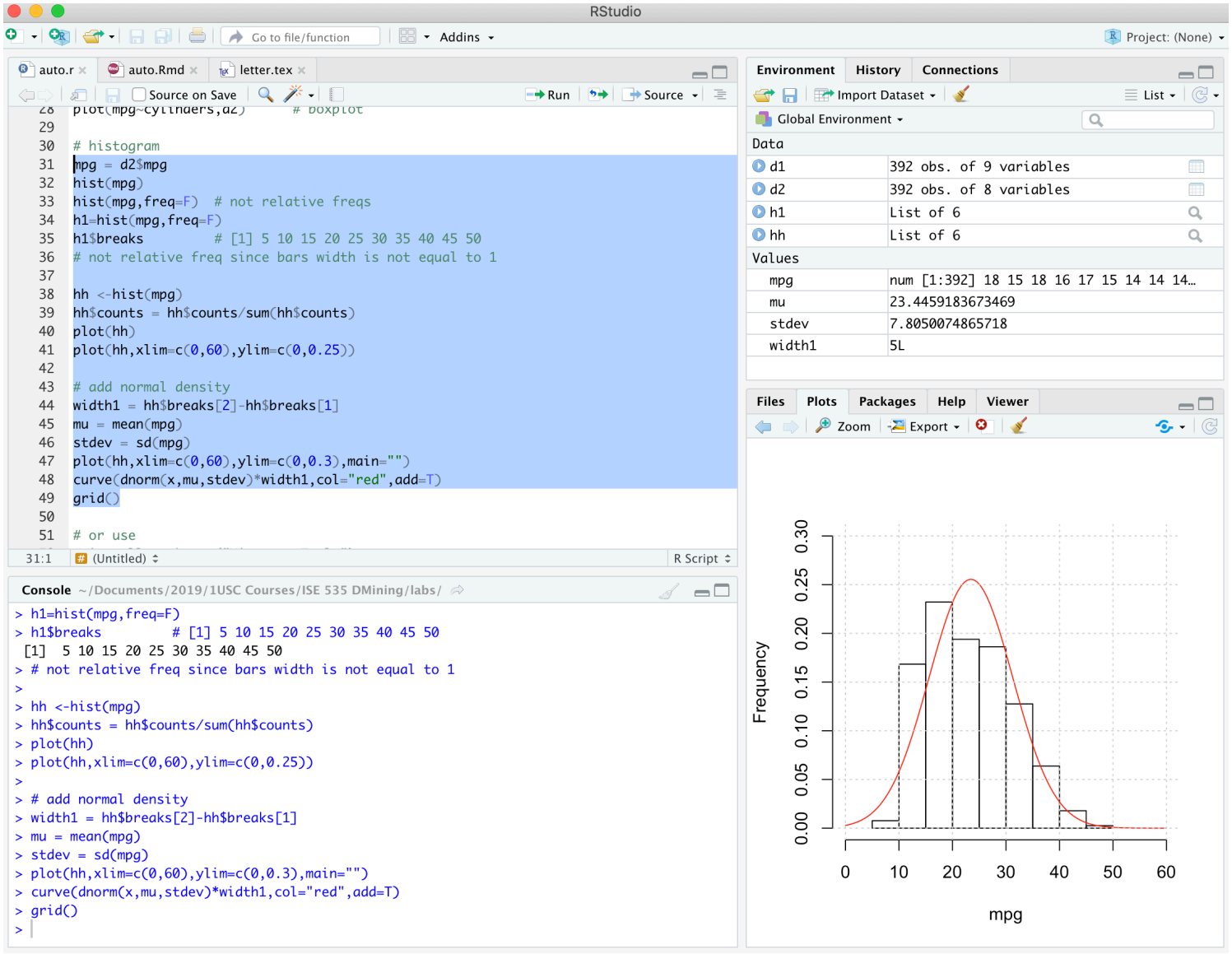
- an Editor
- an Integrated Development Environment (IDE)

Introduction

RStudio is an IDE (Integrated Development Environment)

It includes a text editor, interpreter, graphical display window, and additional features

RStudio


RStudio
auto.r
auto.Rmd
letter.tex
Source on Save
Run
Source
# histogram
mpg = d2$mpg
hist(mpg)
hist(mpg,freq=F) # not relative freqs
h1=hist(mpg,freq=F)
h1$breaks # [1] 5 10 15 20 25 30 35 40 45 50
# not relative freq since bars width is not equal to 1
hh <-hist(mpg)
hh$counts = hh$counts/sum(hh$counts)
plot(hh)
plot(hh,xlim=c(0,60),ylim=c(0,0.25))
# add normal density
width1 = hh$breaks[2]-hh$breaks[1]
mu = mean(mpg)
stdev = sd(mpg)
plot(hh,xlim=c(0,60),ylim=c(0,0.3),main="")
curve(dnorm(x,mu,stdev)*width1,col="red",add=T)
grid()
# or use
31:1
(Untitled)
R Script
Console ~/Documents/2019/1USC Courses/ISE 535 DMining/labs/
> h1=hist(mpg,freq=F)
> h1$breaks # [1] 5 10 15 20 25 30 35 40 45 50
[1] 5 10 15 20 25 30 35 40 45 50
> # not relative freq since bars width is not equal to 1
> hh <-hist(mpg)
> hh$counts = hh$counts/sum(hh$counts)
> plot(hh)
> plot(hh,xlim=c(0,60),ylim=c(0,0.25))
> # add normal density
> width1 = hh$breaks[2]-hh$breaks[1]
> mu = mean(mpg)
> stdev = sd(mpg)
> plot(hh,xlim=c(0,60),ylim=c(0,0.3),main="")
> curve(dnorm(x,mu,stdev)*width1,col="red",add=T)
> grid()
Environment
History
Connections
Import Dataset
List
Global Environment
Data
d1 392 obs. of 9 variables
d2 392 obs. of 8 variables
h1 List of 6
hh List of 6
Values
mpg num [1:392] 18 15 18 16 17 15 14 14 14…
mu 23.4459183673469
stdev 7.8050074865718
width1 5L
Files
Plots
Packages
Help
Viewer
Zoom
Export
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
50
60
mpg

RStudio Notes

- Set the working directory
  - Session
    - Set Working Directory
      - To Source File Location

What is the difference between R and Python?

- R is best for *statistical learning* methods

- Python is best for *machine learning* tools

What is rmarkdown?

Package 'rmarkdown'

December 12, 2019

Type Package

Title Dynamic Documents for R

Version 2.0

Maintainer Yihui Xie <xie@yihui.name>

Description Convert R Markdown documents into a variety of formats.

Depends R (>= 3.0)

Imports tools, utils, knitr (>= 1.22), yaml (>= 2.1.19), htmltools (>= 0.3.5), evaluate (>= 0.13), base64enc, jsonlite, mime, tinytex (>= 0.11), xfun, methods, stringr (>= 1.2.0)

Suggests shiny (>= 0.11), tufte, testthat, digest, dygraphs, tibble, fs, pkgdown, callr (>= 2.0.0)

SystemRequirements pandoc (>= 1.12.3) - http://pandoc.org

URL https://github.com/rstudio/rmarkdown

BugReports https://github.com/rstudio/rmarkdown/issues

Rmarkdown

an R library to create dynamic documents

What is a dynamic document?

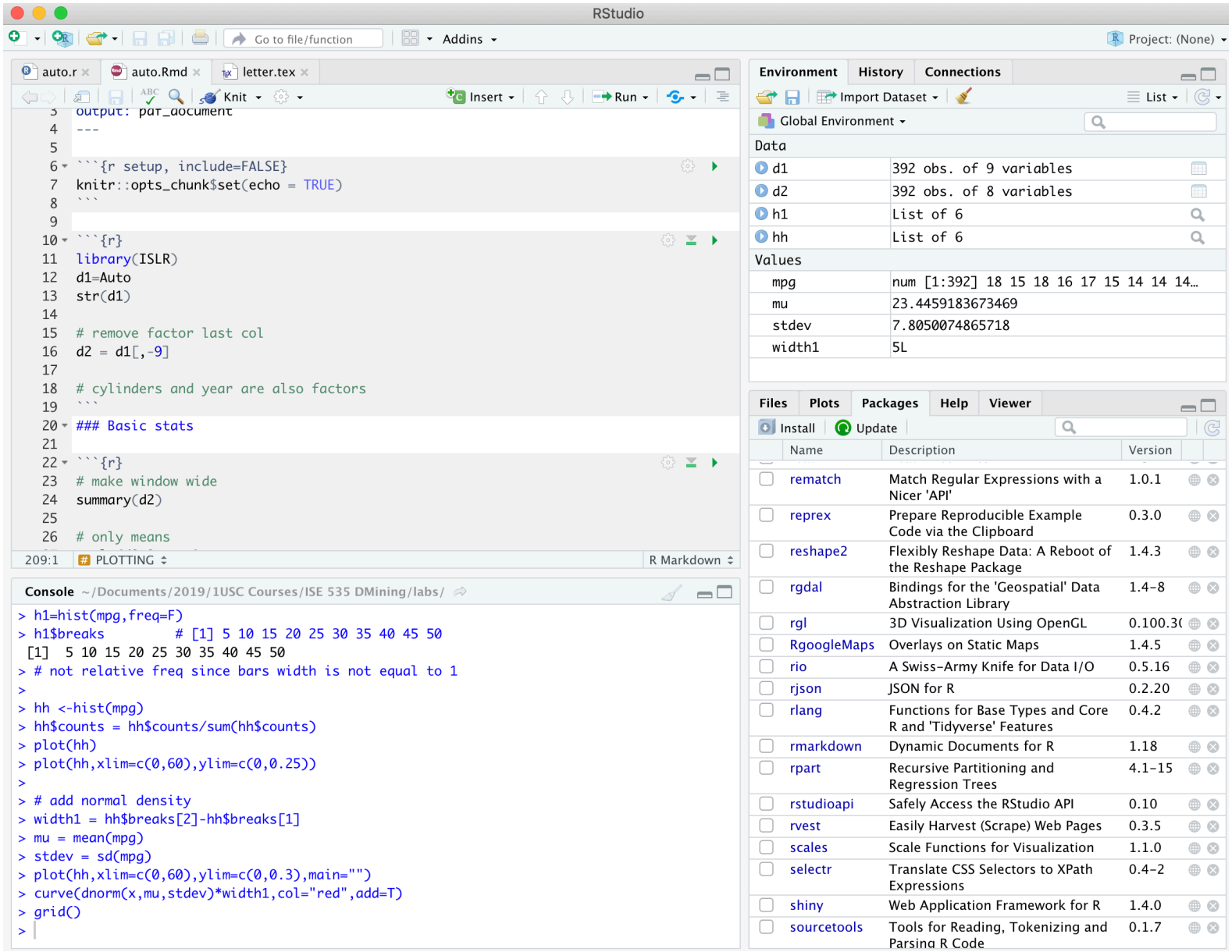
A document that includes text, scripts, output

Using R Markdown

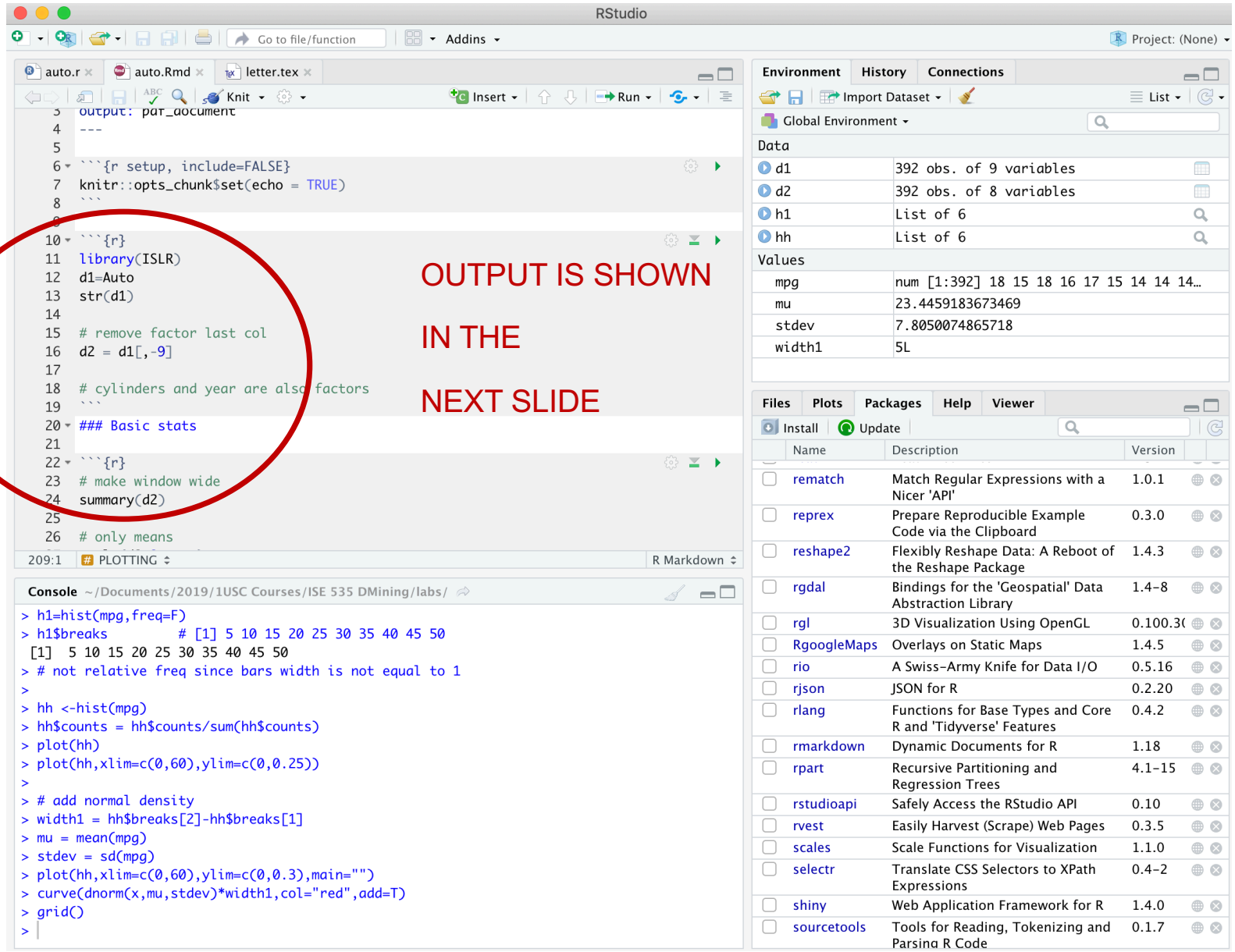
- To install *rmarkdown*
  - install.packages("rmarkdown")
  - tinytex::install_tinytex()
- Most LaTeX commands can be used in .Rmd
- RStudio can be used as compiler of LaTeX scripts

Creating rmarkdown file

- New File > R Markdown
- Save file (will be created as .Rmd)
- Add R code and text portions into the .Rmd file
- Knit the file (generating pdf file)


RStudio
Go to file/function
Addins
Project: (None)
auto.r
auto.Rmd
letter.tex
Knit
Insert
Run
output: pdf_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
library(ISLR)
d1=Auto
str(d1)
# remove factor last col
d2 = d1[,-9]
# cylinders and year are also factors
```
Basic stats
```{r}
# make window wide
summary(d2)
# only means
209:1
PLOTTING
R Markdown
Console ~/Documents/2019/1USC Courses/ISE 535 DMining/labs/
> h1=hist(mpg,freq=F)
> h1$breaks # [1] 5 10 15 20 25 30 35 40 45 50
[1] 5 10 15 20 25 30 35 40 45 50
> # not relative freq since bars width is not equal to 1
>
> hh <-hist(mpg)
> hh$counts = hh$counts/sum(hh$counts)
> plot(hh)
> plot(hh,xlim=c(0,60),ylim=c(0,0.25))
>
> # add normal density
> width1 = hh$breaks[2]-hh$breaks[1]
> mu = mean(mpg)
> stdev = sd(mpg)
> plot(hh,xlim=c(0,60),ylim=c(0,0.3),main="")
> curve(dnorm(x,mu,stdev)*width1,col="red",add=T)
> grid()
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
d1 392 obs. of 9 variables
d2 392 obs. of 8 variables
h1 List of 6
hh List of 6
Values
mpg num [1:392] 18 15 18 16 17 15 14 14 14…
mu 23.4459183673469
stdev 7.8050074865718
width1 5L
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
rematch Match Regular Expressions with a Nicer 'API' 1.0.1
reprex Prepare Reproducible Example Code via the Clipboard 0.3.0
reshape2 Flexibly Reshape Data: A Reboot of the Reshape Package 1.4.3
rgdal Bindings for the 'Geospatial' Data Abstraction Library 1.4-8
rgl 3D Visualization Using OpenGL 0.100.3(
RgoogleMaps Overlays on Static Maps 1.4.5
rio A Swiss-Army Knife for Data I/O 0.5.16
rjson JSON for R 0.2.20
rlang Functions for Base Types and Core R and 'Tidyverse' Features 0.4.2
rmarkdown Dynamic Documents for R 1.18
rpart Recursive Partitioning and Regression Trees 4.1-15
rstudioapi Safely Access the RStudio API 0.10
rvest Easily Harvest (Scrape) Web Pages 0.3.5
scales Scale Functions for Visualization 1.1.0
selectr Translate CSS Selectors to XPath Expressions 0.4-2
shiny Web Application Framework for R 1.4.0
sourcetools Tools for Reading, Tokenizing and Parsing R Code 0.1.7
```


RStudio
Go to file/function
Addins
Project: (None)
auto.r
auto.Rmd
letter.tex
Knit
Insert
Run
output: par_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
library(ISLR)
d1=Auto
str(d1)
# remove factor last col
d2 = d1[,-9]
# cylinders and year are also factors
```
Basic stats
```{r}
# make window wide
summary(d2)
# only means
OUTPUT IS SHOWN
IN THE
NEXT SLIDE
209:1
PLOTTING
R Markdown
Console ~/Documents/2019/1USC Courses/ISE 535 DMining/labs/
> h1=hist(mpg,freq=F)
> h1$breaks # [1] 5 10 15 20 25 30 35 40 45 50
[1] 5 10 15 20 25 30 35 40 45 50
> # not relative freq since bars width is not equal to 1
>
> hh <-hist(mpg)
> hh$counts = hh$counts/sum(hh$counts)
> plot(hh)
> plot(hh,xlim=c(0,60),ylim=c(0,0.25))
>
> # add normal density
> width1 = hh$breaks[2]-hh$breaks[1]
> mu = mean(mpg)
> stdev = sd(mpg)
> plot(hh,xlim=c(0,60),ylim=c(0,0.3),main="")
> curve(dnorm(x,mu,stdev)*width1,col="red",add=T)
> grid()
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
d1 392 obs. of 9 variables
d2 392 obs. of 8 variables
h1 List of 6
hh List of 6
Values
mpg num [1:392] 18 15 18 16 17 15 14 14 14…
mu 23.4459183673469
stdev 7.8050074865718
width1 5L
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
rematch Match Regular Expressions with a Nicer 'API' 1.0.1
reprex Prepare Reproducible Example Code via the Clipboard 0.3.0
reshape2 Flexibly Reshape Data: A Reboot of the Reshape Package 1.4.3
rgdal Bindings for the 'Geospatial' Data Abstraction Library 1.4-8
rgl 3D Visualization Using OpenGL 0.100.3(
RgoogleMaps Overlays on Static Maps 1.4.5
rio A Swiss-Army Knife for Data I/O 0.5.16
rjson JSON for R 0.2.20
rlang Functions for Base Types and Core R and 'Tidyverse' Features 0.4.2
rmarkdown Dynamic Documents for R 1.18
rpart Recursive Partitioning and Regression Trees 4.1-15
rstudioapi Safely Access the RStudio API 0.10
rvest Easily Harvest (Scrape) Web Pages 0.3.5
scales Scale Functions for Visualization 1.1.0
selectr Translate CSS Selectors to XPath Expressions 0.4-2
shiny Web Application Framework for R 1.4.0
sourcetools Tools for Reading, Tokenizing and Parsing R Code 0.1.7
```

```
library(ISLR)
d1=Auto
str(d1)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg          : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders    : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower   : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight       : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year         : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name         : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14
```

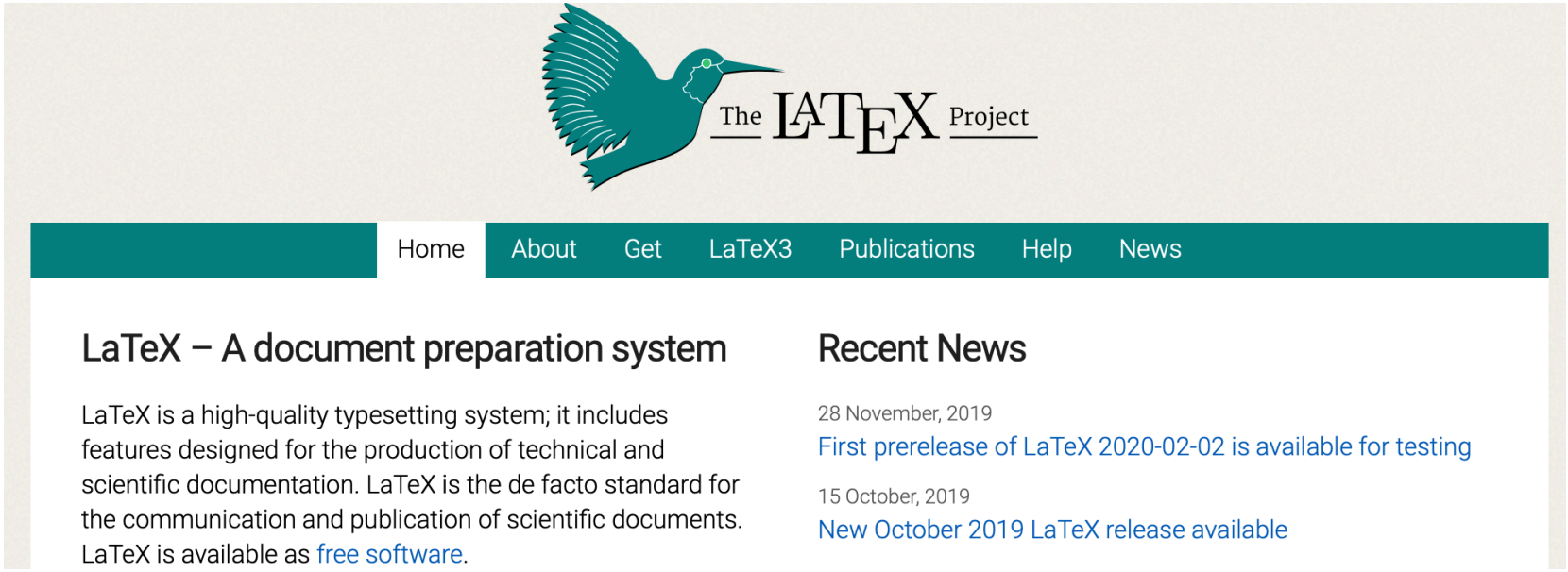
```
# remove factor last col
d2 = d1[,-9]

# cylinders and year are also factors
```

Basic stats

```
# make window wide
summary(d2)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
```

R Markdown requires LaTeX



LaTeX is used –alone- to create documents

R Markdown notes

- tinytex::install_tinytex() will install latex in RStudio
- Most LaTeX commands can be used in .Rmd
- RStudio can be used as compiler of LaTeX scripts

ISE 599 Data Mining

- We will learn how to create R Markdown documents
- Homework must be submitted as R Markdown documents, in class