
Getting started with R and R Studio a 'Webin-R'

Ana Morales-Gómez
UK Data Service



NCRM-UKDS
November 2019



Overview of this webinar

- ✓ Introduction
 - What is R and R Studio?
 - How to get R and R Studio? (downloading and installing)
 - R Studio environment
- ✓ Getting Started
- ✓ Data types and Structures
- ✓ Using data

Introduction: What are R and R Studio



- R is a statistical programming language
- Open source
- Free
- Available for Windows, Macintosh, and Linux.
- Huge community of users and developers
- Scripting language, i.e. uses code

- **Integrated Development Environment or IDE**
- All of R goodies, plus
- User friendly interface
- Need R installed

Download and installing



[Home]

Download

CRAN

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

<https://www.r-project.org/>

Open Source Edition

Overview

- Access RStudio locally
- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors quickly
- Extensive package development tools

Support Community forums only

License AGPL v3

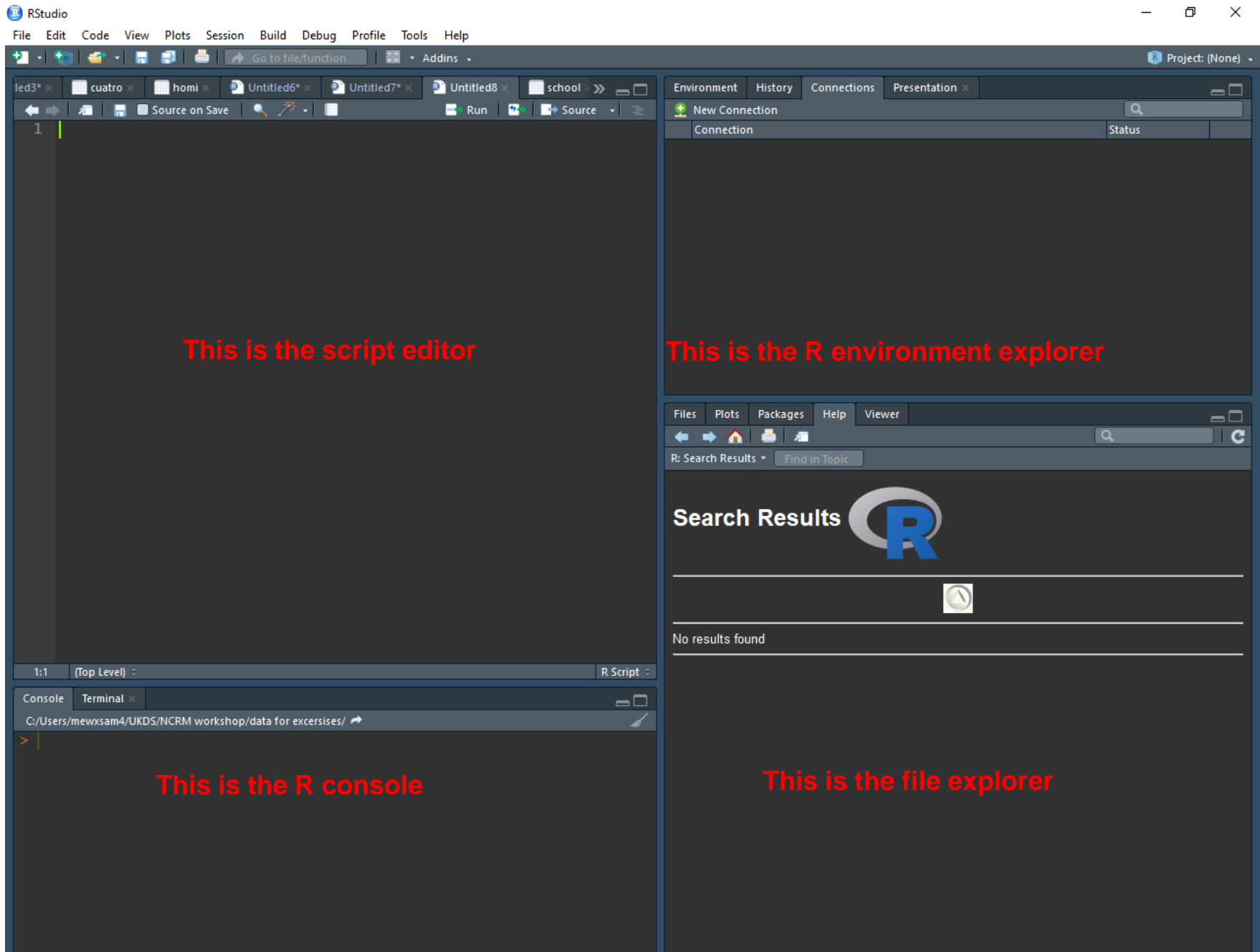
Pricing Free

[DOWNLOAD RSTUDIO DESKTOP](#)



<https://www.rstudio.com/products/rstudio/download/>

R Studio Interface



This is the script editor

This is the R environment explorer

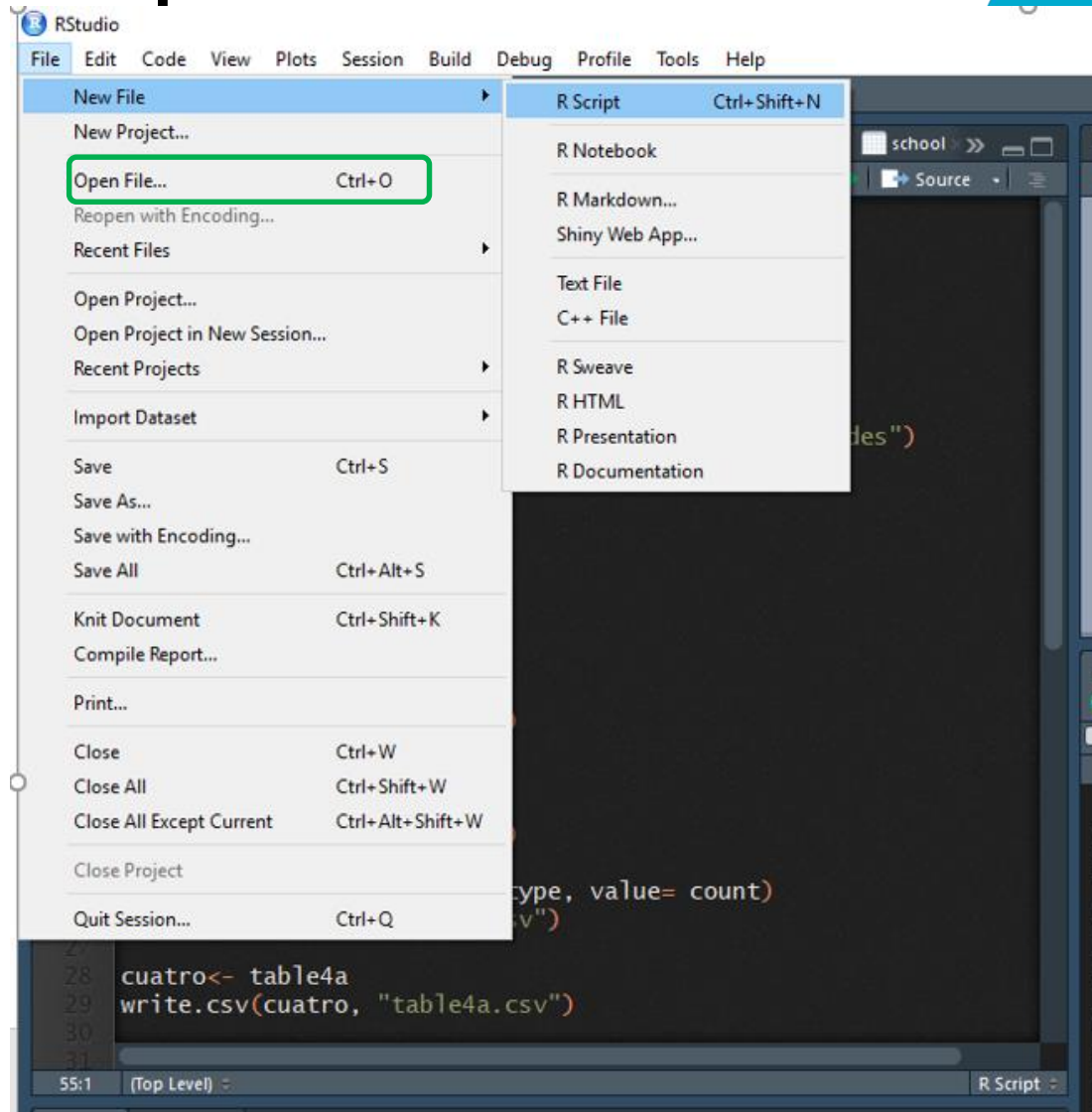
This is the R console

This is the file explorer

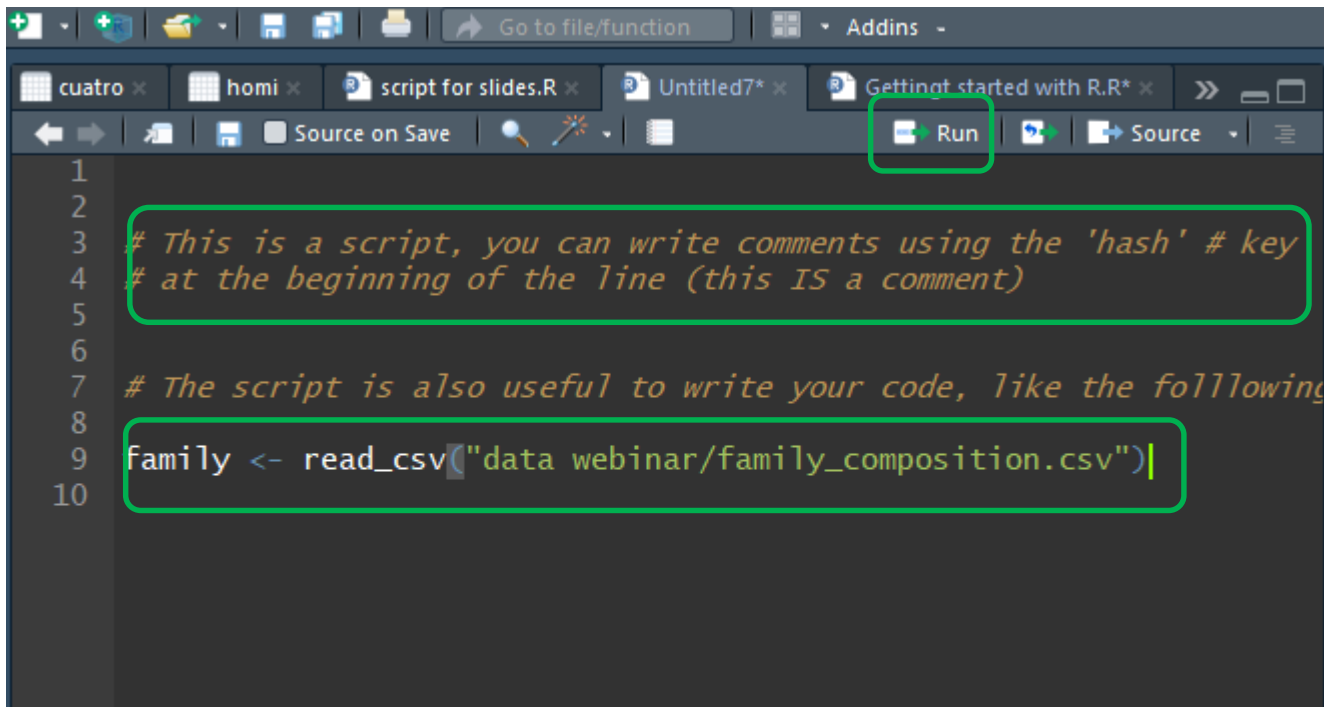
Getting started with R: Scripts

- ✓ Scripts are used to save our work and analyses
 - Can be stored as R script or Notepad
 - Can be opened again in later sessions
 - Can be copied and modified
 - Can be shared

```
4
5
6 # Create folders in your wd
7 dir.create("dataSlides")
8
9 setwd("C:/Users/mewxsam4/UKDS/NCRM workshop/dataSlides")
10
11
12 library(haven)
13 library(tidyverse)
14
15 #tidying data
16
17 uno<- table1
18
19 write.csv(uno, "table1.csv")
20
21 dos<-table2
22 table2
23 write.csv(dos, "table2.csv")
24
25 dostidy<- spread(dos, key= type, value= count)
26
```



Scripts

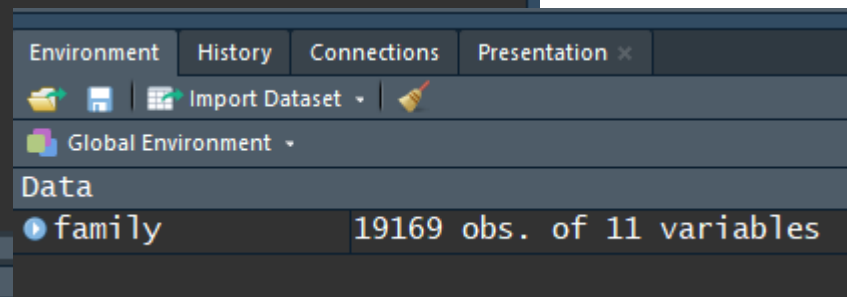


The screenshot shows an R script editor with a dark background. The top toolbar includes a 'Run' button (a play icon) which is highlighted with a green box. Below the toolbar, the script content is displayed with line numbers 1 through 10. Lines 3 and 4 contain comments starting with '#'. Lines 7 and 8 also contain comments. Line 9 contains the R code `family <- read_csv("data webinar/family_composition.csv")`, which is highlighted with a green box. The status bar at the bottom left shows the time 9:58 and '(Top Level)'.

```
1
2
3 # This is a script, you can write comments using the 'hash' # key
4 # at the beginning of the line (this IS a comment)
5
6
7 # The script is also useful to write your code, like the following
8
9 family <- read_csv("data webinar/family_composition.csv")
10
```

You can select a code and press 'Run'

Or, click/select on the line of the code and press:
Ctrl + Enter (windows)
Command+Alt+R (Mac)



The screenshot shows the R Environment pane with tabs for Environment, History, Connections, and Presentation. The Environment tab is active, showing a 'Global Environment' section and a 'Data' section. Under 'Data', there is a table with one row: 'family' with '19169 obs. of 11 variables'.

Environment	History	Connections	Presentation ×
Import Dataset			
Global Environment			
Data			
family	19169 obs. of 11 variables		

Working directory...

- ✓ Tells R where our data is saved in our PC, laptops, external drive.
- ✓ Tells R where to save our new analyses and figures
- ✓ Code to set the working directory:

```
> setwd("your/folder/path")
```

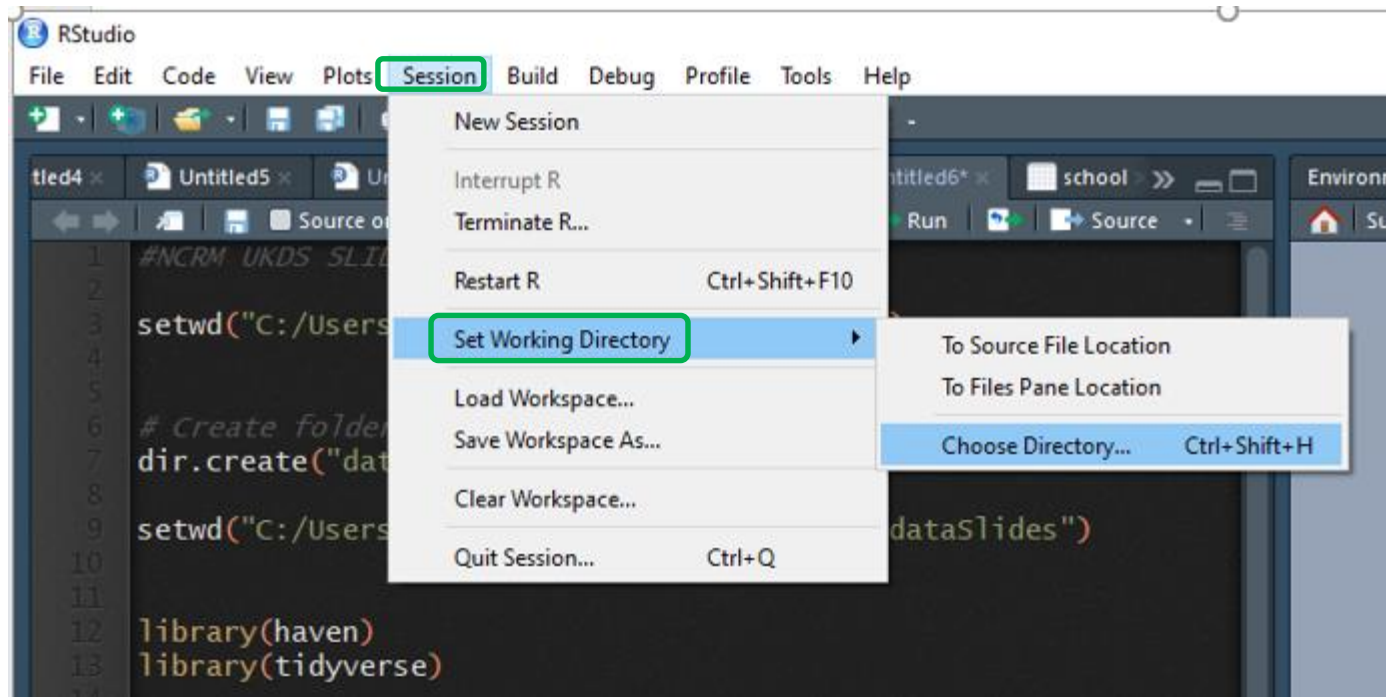
To check where the working directory (wd) is:

```
> getwd()
```

- ✓ OR...



Working directory



```
Console Terminal R Markdown
C:/Users/mewxsam4/UKDS/NCRM workshop/ ↗
>
>
>
> getwd()
[1] "C:/Users/mewxsam4/UKDS/NCRM workshop"
```

Packages

- ✓ Collection of R functions, compiled in a defined format
- ✓ Set of basic pre-installed operations
- ✓ R needs packages to do certain tasks
 - haven: For importing datasets in other formats (SPSS, Stata, SAS).
 - ggplot2: For producing graphs
 - tmap: For producing maps

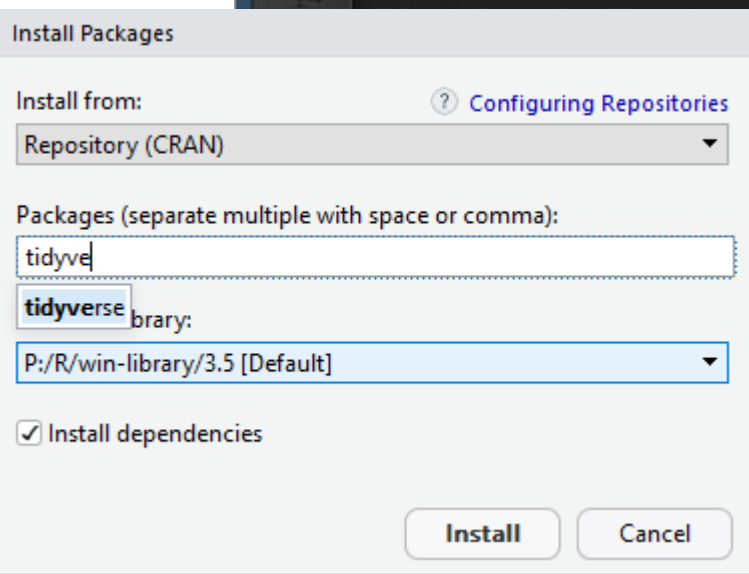
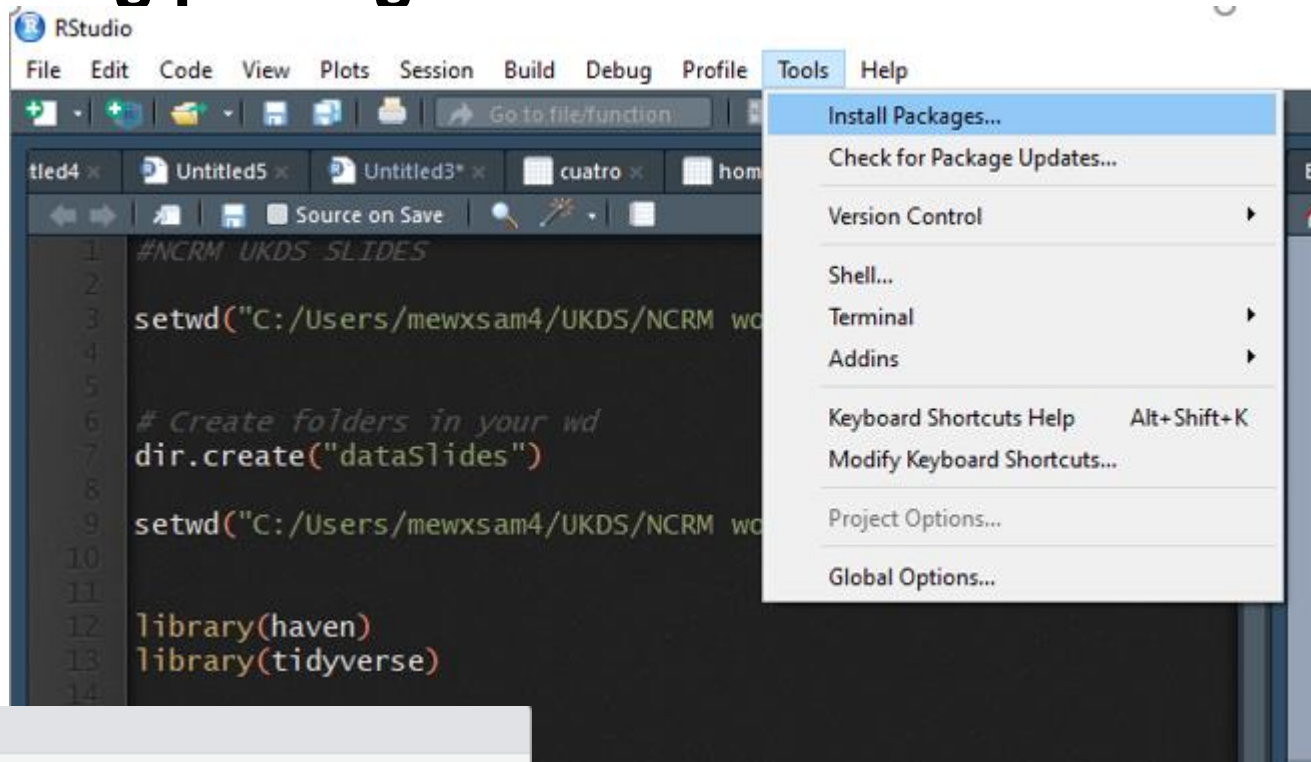
✓ Code

- > `install.packages("haven")`
- > `install.packages("haven", "ggplot2")`

OR...



Installing packages



```
> install.packages('tidyverse')
Installing package into 'P:/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/tidyverse_1
.2.1.zip'
Content type 'application/zip' length 92570 bytes (90 KB)
downloaded 90 KB
```

Loading packages

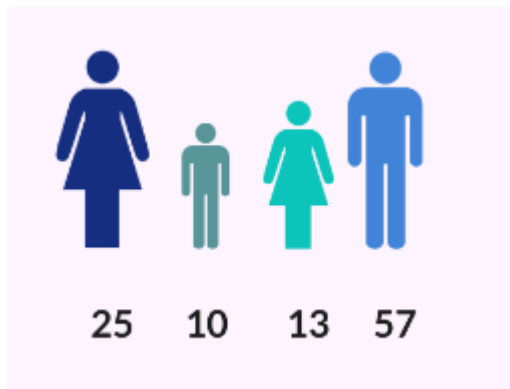
```
> library(tidyverse)
-- Attaching packages ----- tidyverse
rse 1.2.1 --
v ggplot2 2.2.1      v purrr  0.2.4
v tibble  1.4.2      v dplyr  0.7.6
v tidyr   0.8.0      v stringr 1.4.0
v readr   1.1.1      v forcats 0.3.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
Warning messages:
1: package 'tidyverse' was built under R version 3.5.3
2: package 'stringr' was built under R version 3.5.3
> |
```

- ✓ Each package needs to be loaded every time you start a new R session
- ✓ Only load the package that you need to use
- ✓ Can be done at any time
- ✓ Indicate in the script the packages used

Data types and data Structures

✓ Data types

- character
- numeric (real or decimal)
- integer
- logical



✓ Structures

- Vectors (variables)
- factors
- list
- matrix
- data frame



Variables

- Variables are objects in R that store values;
- The “<-” tells R to take the number to the right of the symbol and store it in a variable whose name is given on the left.

```
> 3
[1] 3
> a <- 3
> a
[1] 3
> |
```

```
> b <- 5
> c <- 9
>
> b*c
[1] 45
> b*c/a
[1] 15
|
```

```
> d <- b*c/a
> d
[1] 15
```

Vectors

- ✓ vectors are 'a single entity consisting of a collection of things'
 - a in this example is a vector of length 1
- ✓ Longer vectors can be created by *concatenating* 'c' values
- ✓ There are several types of vectors such as character vectors, numeric, logical, etc.
 - For example: The typical variable age in a dataset is a 'vector'

```
> 3
[1] 3
> a <- 3
> a
[1] 3
> |
```

```
> v <- c(a, b, c)
> v
[1] 3 5 9
> v1 <- c(3, 5, 9)
> v1
[1] 3 5 9
```

Data frames and Tibbles

- ✓ Data frames are the '*de facto*' data structure for tabular data.
- ✓ Tibbles *are* data frames, but with some tweaks.
 - Designed specially to work well within the 'tidyverse' package

```
> as.data.frame(table1)
  country year cases population
1 Afghanistan 1999    745  19987071
2 Afghanistan 2000   2666  20595360
3    Brazil 1999  37737  172006362
4    Brazil 2000  80488  174504898
5    China 1999 212258 1272915272
6    China 2000 213766 1280428583
```

```
> table1
# A tibble: 6 x 4
  country      year cases population
  <chr>      <int> <int>    <int>
1 Afghanistan  1999     745  19987071
2 Afghanistan  2000    2666  20595360
3    Brazil    1999   37737  172006362
4    Brazil    2000   80488  174504898
5    China     1999  212258 1272915272
6    China     2000  213766 1280428583
```

Reference: R for data science chapter 10

<https://r4ds.had.co.nz/tibbles.html>

Importing data

- ✓ Get the appropriate package:

- haven
- foreign
- readr

- ✓ Use the right function:

- Examples using functions from 'haven' and 'readr' package

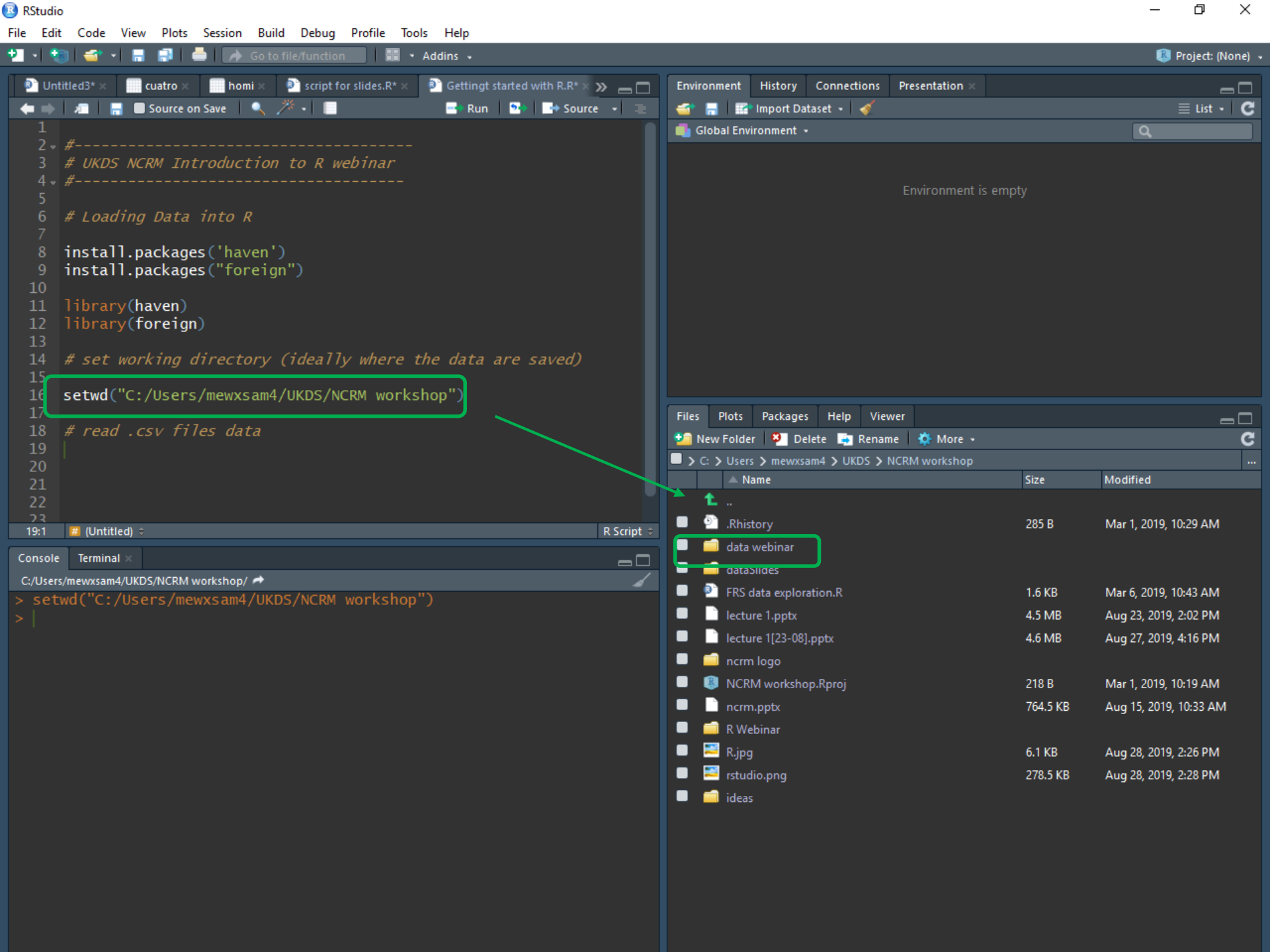
Csv files: `read_csv("mydata.csv")`

Stata files: `read_dta("mydata.dta")`

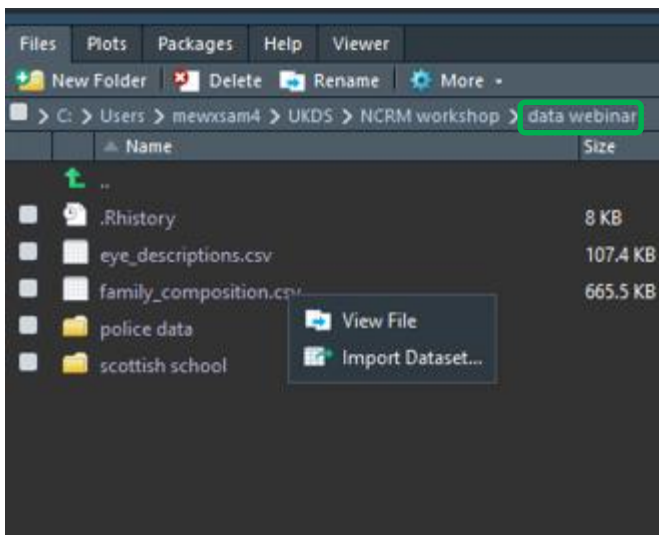
SPSS files: `read_sav("mydata.sav")`

- ✓ Give your data a name!: `census<- read_dta("mydata.dta")`





Importing data, the easy way



Double click on the folder where the data is

Click on the data we want to import: family_composition.csv

Click on 'import dataset'...

Reference: R for data science chapter 11
<https://r4ds.had.co.nz/data-import.html>

Import Text Data

File/Url:

C:/Users/mewxsam4/UKDS/NCRM workshop/data webinar/family_composition.csv

Update

Data Preview:

user_id <small>(integer)</small>	sex <small>(character)</small>	age <small>(double)</small>	momage <small>(integer)</small>	dadage <small>(integer)</small>	oldbro <small>(integer)</small>	oldsis <small>(integer)</small>	youngbro <small>(integer)</small>	youngsis <small>(integer)</small>	twinbro <small>(integer)</small>	twinsis <small>(integer)</small>
8	male	38.1	25	27	0	0	0	1	0	0
67	female	19.7	29	31	1	0	0	1	0	0
98	female	19.4	NA	NA	1	0	0	1	0	0
103	female	20.6	NA	NA	2	0	0	0	0	0
164	female	20.3	24	NA	0	0	0	0	0	0
233	female	19.3	NA	NA	0	2	0	0	0	0
235	male	18.7	NA	NA	0	0	1	0	0	0
253	female	19.5	24	25	0	0	1	0	0	0
256	female	19.7	NA	NA	1	1	0	0	0	0
271	female	24.5	21	22	0	0	2	2	0	0
298	female	17.7	28	NA	0	0	1	0	0	0
332	male	19.6	NA	NA	1	0	0	0	0	0
426	male	19.2	NA	NA	0	0	2	0	0	0
429	female	19.8	NA	NA	1	4	0	0	0	0
434	male	18.8	NA	NA	1	0	0	0	0	0
436	female	22.1	NA	NA	2	0	2	0	0	0
450	female	19.2	NA	NA	0	0	0	1	0	0
452	female	19.4	NA	NA	1	0	1	1	0	0
474	male	49.4	26	30	0	2	1	0	0	0

Previewing first 50 entries.

Import Options:

Name: First Row as Names
 Skip: Trim Spaces
 Open Data Viewer
 Delimiter: Escape:
 Quotes: Comment:
 Locale: NA:

Code Preview:

```
library(readr)
family_composition <- read_csv("data
webinar/family_composition.csv")
view(family_composition)
```

? Reading rectangular data using readr

Import Cancel

Go to file/function Addins

omi x script for slides.R x Getting started with R.R x family_composition x Untitled >>

Filter

	user_id	sex	age	momage	dadage	oldbro	oldsis	youngbro	youngsis	twin
1	8	male	38.1	25	27	0	0	0	0	1
2	67	female	19.7	29	31	1	0	0	0	1
3	98	female	19.4	NA	NA	1	0	0	0	1
4	103	female	20.6	NA	NA	2	0	0	0	0
5	164	female	20.3	24	NA	0	0	0	0	0
6	233	female	19.3	NA	NA	0	2	0	0	0
7	235	male	18.7	NA	NA	0	0	1	0	0
8	253	female	19.5	24	25	0	0	1	0	0
9	256	female	19.7	NA	NA	1	1	0	0	0
10	271	female	24.5	21	22	0	0	2	2	2
11	298	female	17.7	28	NA	0	0	1	0	0
12	332	male	19.6	NA	NA	1	0	0	0	0
13	426	male	19.2	NA	NA	0	0	2	0	0
14	429	female	19.8	NA	NA	1	4	0	0	0
15	434	male	18.8	NA	NA	1	0	0	0	0
16	435	female	22.1	NA	NA	2	0	2	0	0

Showing 1 to 17 of 19,169 entries

Console Terminal x

```
C:/Users/mewxsam4/UKDS/NCRM workshop/
> setwd("C:/Users/mewxsam4/UKDS/NCRM workshop")
> library(readr)
> family_composition <- read_csv("data webinar/family_composition.csv")
Parsed with column specification:
cols(
  user_id = col_integer(),
  sex = col_character(),
  age = col_double(),
  momage = col_integer(),
  dadage = col_integer(),
  oldbro = col_integer(),
  oldsis = col_integer(),
  youngbro = col_integer(),
  youngsis = col_integer(),
  twinbro = col_integer(),
  twinsis = col_integer()
)
> View(family_composition)
> |
```

Environment History Connections Presentation x

Import Dataset

Global Environment

Data

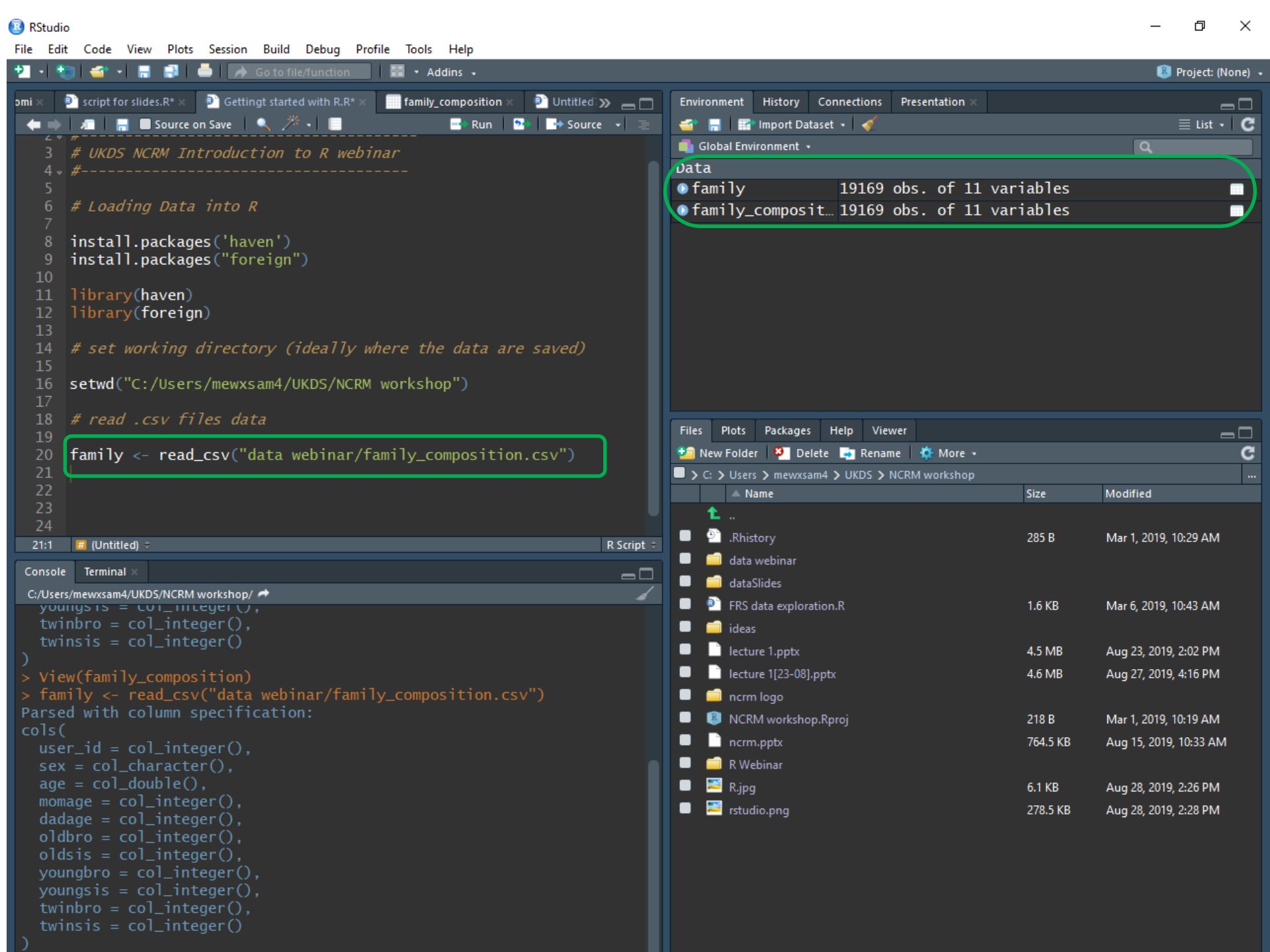
family_composit... 19169 obs. of 11 variables

Files Plots Packages Help Viewer

New Folder Delete Rename More

C:\Users\mewxsam4\UKDS\NCRM workshop\data webinar

Name	Size	Modified
..		
.Rhistory	8 KB	Aug 28, 2019, 4:11 PM
eye_descriptions.csv	107.4 KB	Aug 28, 2019, 9:44 AM
family_composition.csv	665.5 KB	Aug 28, 2019, 9:49 AM
police_data		
scottish school		



```
2 #  
3 # UKDS NCRM Introduction to R webinar  
4 #-----  
5  
6 # Loading Data into R  
7  
8 install.packages('haven')  
9 install.packages("foreign")  
10  
11 library(haven)  
12 library(foreign)  
13  
14 # set working directory (ideally where the data are saved)  
15  
16 setwd("C:/Users/mewxsam4/UKDS/NCRM workshop")  
17  
18 # read .csv files data  
19  
20 family <- read_csv("data webinar/family_composition.csv")  
21  
22  
23  
24
```

```
Console Terminal  
C:/Users/mewxsam4/UKDS/NCRM workshop/  
youngsis = col_integer(),  
twinbro = col_integer(),  
twinsis = col_integer()  
)  
> View(family_composition)  
> family <- read_csv("data webinar/family_composition.csv")  
Parsed with column specification:  
cols(  
  user_id = col_integer(),  
  sex = col_character(),  
  age = col_double(),  
  momage = col_integer(),  
  dadage = col_integer(),  
  oldbro = col_integer(),  
  oldsis = col_integer(),  
  youngbro = col_integer(),  
  youngsis = col_integer(),  
  twinbro = col_integer(),  
  twinsis = col_integer()  
)
```

Environment History Connections Presentation

Global Environment

Data

- family 19169 obs. of 11 variables
- family_composit... 19169 obs. of 11 variables

Files Plots Packages Help Viewer

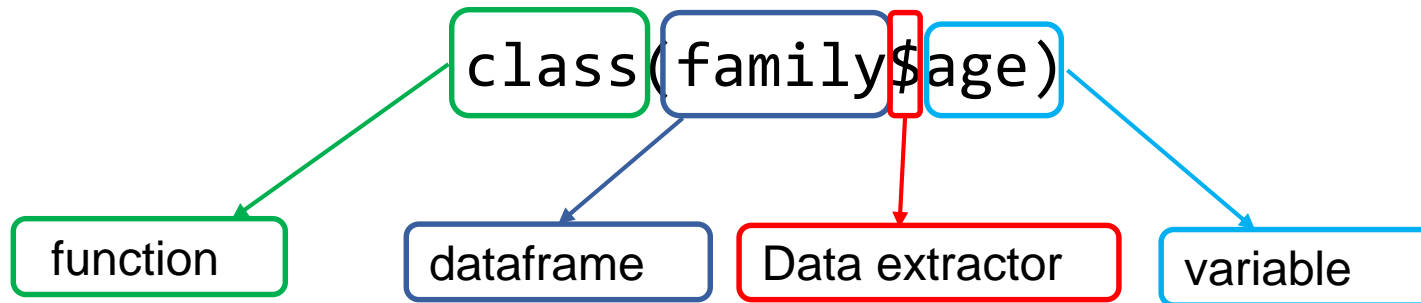
New Folder Delete Rename More

C:\Users\mewxsam4\UKDS\NCRM workshop

Name	Size	Modified
..		
.Rhistory	285 B	Mar 1, 2019, 10:29 AM
data webinar		
dataSlides		
FRS data exploration.R	1.6 KB	Mar 6, 2019, 10:43 AM
ideas		
lecture 1.pptx	4.5 MB	Aug 23, 2019, 2:02 PM
lecture 1[23-08].pptx	4.6 MB	Aug 27, 2019, 4:16 PM
nCRM logo		
NCRM workshop.Rproj	218 B	Mar 1, 2019, 10:19 AM
nCRM.pptx	764.5 KB	Aug 15, 2019, 10:33 AM
R Webinar		
R.jpg	6.1 KB	Aug 28, 2019, 2:26 PM
rstudio.png	278.5 KB	Aug 28, 2019, 2:28 PM

Using data in R

- To perform operations on specific variables, we need to specify the data frame and the variable: `class(family$age)`



```
Console Terminal x R Markdown x
C:/Users/mewxsam4/UKDS/NCRM workshop
>
>
> class(family$age)
[1] "numeric"
> |
```



Demo

Recap getting started with R

- First, tell R where your data is; i.e. set your **working directory**

- Second, install/load the required **package(s)**

```
install.packages(ggplot2)
library(ggplot2)
```

- Third, **Import the data**

```
Csv files:    read_csv("mydata.csv")
```

```
Stata files:  read_dta("mydata.dta")
```

```
SPSS files:  read_sav("mydata.sav")
```

Give your data a name!: **census<- read_dta("mydata.dta")**

- Remember

- R is case sensitive, be careful with spaces and capitals/lower case
- Choose an informative and easy to type name for your data
 - You will need to write it a lot while you analyse!

Where to go if you are stuck

- Trial and error (actually errors... and lots of them!)
- Search code online:
 - Wickham and Grolemund, 2016. **R For Data Science**. Available: <https://r4ds.had.co.nz/>
 - Quick R: <http://www.statmethods.net/>
 - <http://www.ats.ucla.edu/stat/r/>
 - <http://stackoverflow.com/>
 - <https://stats.stackexchange.com/>
 - <https://github.com/trending/r>
 - <http://www.cookbook-r.com/>
 - See also the swirl R tutorial on the web <http://swirlstats.com>
 - Or... simply google your questions
- Copy code, modify it if necessary and run it
- Repeat

Questions

Ana Morales-Gomez

ana.morales@manchester.ac.uk

To follow UK Data Service on Twitter:
[@UKDataService](https://twitter.com/UKDataService)

