

R basics

R for Stata Users

Luiza Andrade, Leonardo Viotti & Rob Marty

November-December 2018



Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

These training sessions will offer a quick introduction to R, its amazing features and why it is so much better than Stata (and Excel).

This first session will present the basic concepts you will need to use R.

The next sessions will include:

- **Exporting summary statistics and regression tables**
- **Creating awesome-looking plots**
- **Writing reproducible and readable code**
- **Creating awesome-looking maps**
- **Cleaning data**

For the most recent versions of these trainings, visit the R-training GitHub repo at <https://github.com/worldbank/dime-r-training>

(currently under the 'development' branch)

Why not using Excel?

- Reproducibility of your results.
 - When you add new lines to your data, you don't have to change your analysis.
 - You can re-use your code (or anyone else's) with small changes with entirely different data.
 - Someone else (or your future self) can read your code and see every step of the analysis.
- R is much faster and can do MUCH more.
- R is more forgiving with errors:
 - When you open a sheet in R, you create a virtual copy. So you never change the original file.
 - Data and analysis are separate, so it is easier to see what's wrong.

Some advantages of R over Stata:

- It is less specialized:
 - More flexibility when programming.
 - Many more functionalities.
- Much broader network of users:
 - More resources online, which makes using Google a lot easier. You'll never want to see Statalist again in your life.
 - Development of new features and bug fixes happens faster.
- It is way cooler.

Some possible disadvantages of R:

- Higher cost of entry than Stata.
 - That doesn't mean that the learning curve is steeper all the way up!
- Stata is more specialized:
 - Certain common tasks are simpler in Stata.
- Stata has wider adoption among micro-econometricians.
 - Network externalities in your work environment.
 - Development of new specialized techniques and tools could happen faster (e.g. *ietoolkit*).

Here are some other advantages:

- R is a free and open source software!
- It allows you to have several data sets open simultaneously.
- It can run complex Geographic Information System (GIS) analyses.
- You can use it for web scrapping.
- You can run machine learning algorithms with it.
- You can create complex Markdown documents. This presentation, for example, is entirely done in RStudio.
- You can create interactive dashboards and online applications with the Shiny package.

What about Python?

- Python is even more flexible and has more users than R. So, why should I bother to learn R?
- Despite being super popular for data science, Python has fewer libraries developed for econometrics.
- Python still cannot do everything Stata does without some trouble, R can.
- R and Python are very similar, specially if your background is in Stata.

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

Getting started

This training requires that you have R installed in your computer:

Installation

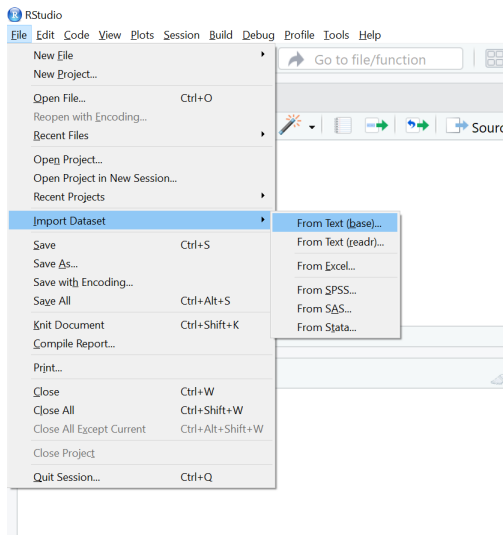
- Please visit (<https://cran.r-project.org>) and select a Comprehensive R Archive Network (CRAN) mirror close to you.
- If you're in the US, you can directly visit the mirror at Berkley university at (<https://cran.cnr.berkeley.edu>).
- we also strongly suggest installing R studio. You can get it in (<https://www.rstudio.com/>), but you need to install R first.

Let's start by loading the data set we'll be using:

Exercise 1: Import data

- 1 In RStudio, go to File > Import Dataset > From Text (Base) and open the `whr_panel.csv` file.
 - Depending on your Rstudio version, it might be File > Import Dataset > From CSV
- 2 The file should be in `GitHub/dime-r-training/DataWork/DataSets/Final`
- 3 Change the name to `whr` on the import window

Getting started



Getting started

Import Dataset

Name

whrf

Encoding

Automatic

Heading

☒ Yes ☐ No

Row names

Automatic

Separator

Comma

Decimal

Period

Quote

Double quote (")

Comment

None

na.strings

NA

☒ Strings as factors

Input File

"country", "region", "year", "happy_rank", "happy_score", "gdp_pc"
"Norway", "Western Europe", 2017, 1, 7.53700017929077, 1.616463184
"Denmark", "Western Europe", 2017, 2, 7.52199983596802, 1.48238301
"Iceland", "Western Europe", 2017, 3, 7.50400018692017, 1.48063302
"Switzerland", "Western Europe", 2017, 4, 7.49399995803833, 1.5645
"Finland", "Western Europe", 2017, 5, 7.4689998626709, 1.443571925
"Netherlands", "Western Europe", 2017, 6, 7.3769998550415, 1.50394
"Canada", "North America", 2017, 7, 7.31599998474121, 1.4792044162
"New Zealand", "Australia and New Zealand", 2017, 8, 7.3140001296
"Sweden", "Western Europe", 2017, 9, 7.28399991989136, 1.494387265
"Australia", "Australia and New Zealand", 2017, 10, 7.283999991989
"Israel", "Middle East and Northern Africa", 2017, 11, 7.21299982
"Costa Rica", "Latin America and Caribbean", 2017, 12, 7.07899999
"Austria", "Western Europe", 2017, 13, 7.00600004196167, 1.487097
"United States", "North America", 2017, 14, 6.99300003051758, 1.54

< >

Data Frame

country	region	year	happy
Norway	Western Europe	2017	1
Denmark	Western Europe	2017	2
Iceland	Western Europe	2017	3
Switzerland	Western Europe	2017	4
Finland	Western Europe	2017	5
Netherlands	Western Europe	2017	6
Canada	North America	2017	7
New Zealand	Australia and New Zealand	2017	8
Sweden	Western Europe	2017	9
Australia	Australia and New Zealand	2017	10
Israel	Middle East and Northern Africa	2017	11
Costa Rica	Latin America and Caribbean	2017	12
Austria	Western Europe	2017	13
United States	North America	2017	14

< >

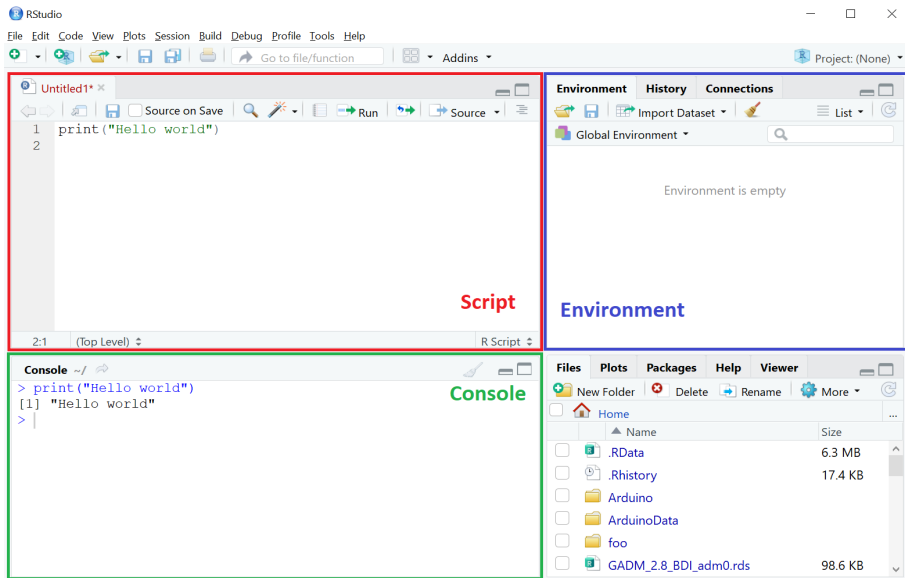
Import

Cancel

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface**
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

RStudio interface



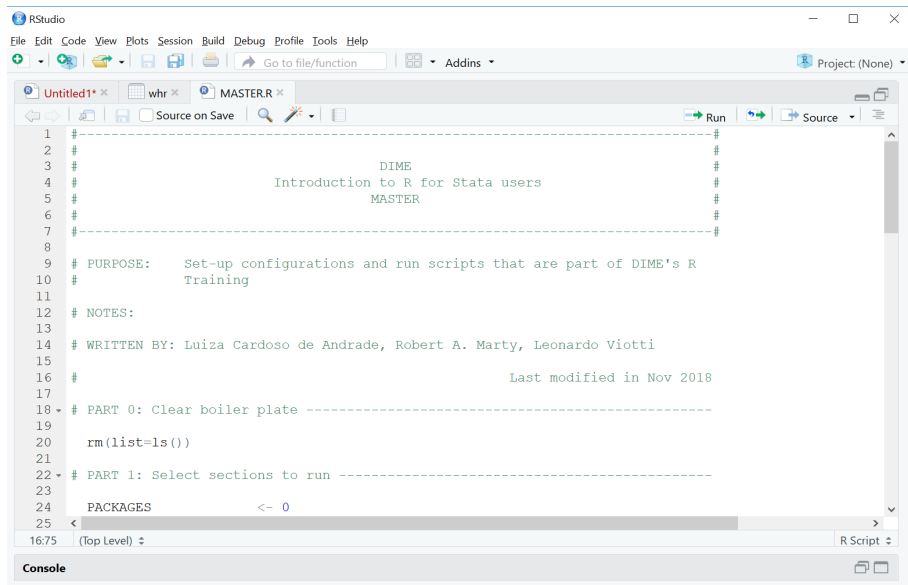
RStudio interface

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for saving, opening, and navigating files. The main workspace is divided into four panes:

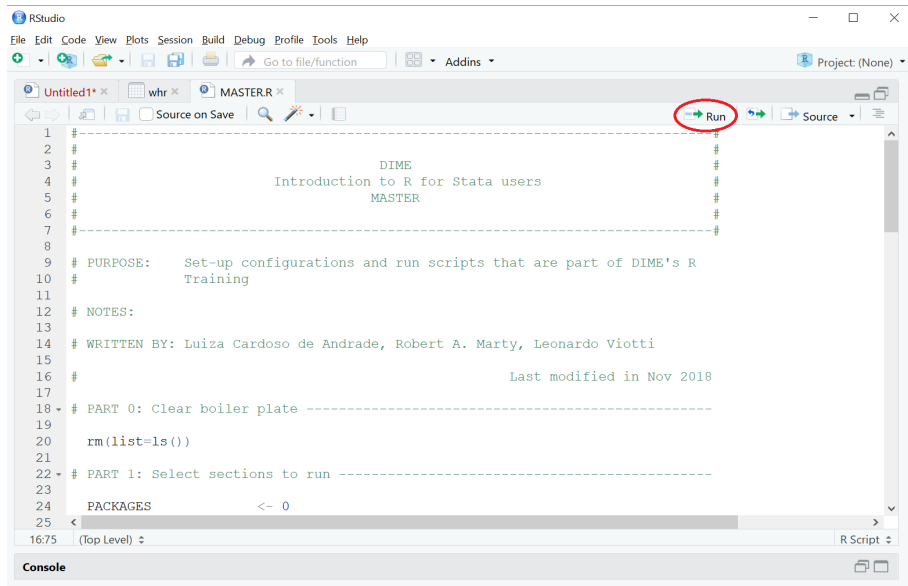
- Source**: Shows a script named 'Untitled1' with a tab for 'whr'. The script contains the following code:

```
> whr <- read.csv("~/GitHub/dime-r-training/DataWork/DataSets/Final/whr_panel.csv")
> View(whr)
> |
```
- Environment**: Displays the loaded object 'whr' with the description '469 obs. of 12 variables'. This entry is circled in red.
- Data Viewer**: Shows a preview of the CSV data with columns: country, region, year, happy_rank, and happy_score. The first seven rows are visible, showing data for Norway, Denmark, Iceland, Switzerland, Finland, Netherlands, and Canada. A scroll bar on the right indicates 469 total entries.
- Files**: Shows the file explorer with a list of files in the current directory, including .RData, .Rhistory, Arduino, ArduinoData, foo, and GADM_2.8_BDI_adm0.rds.

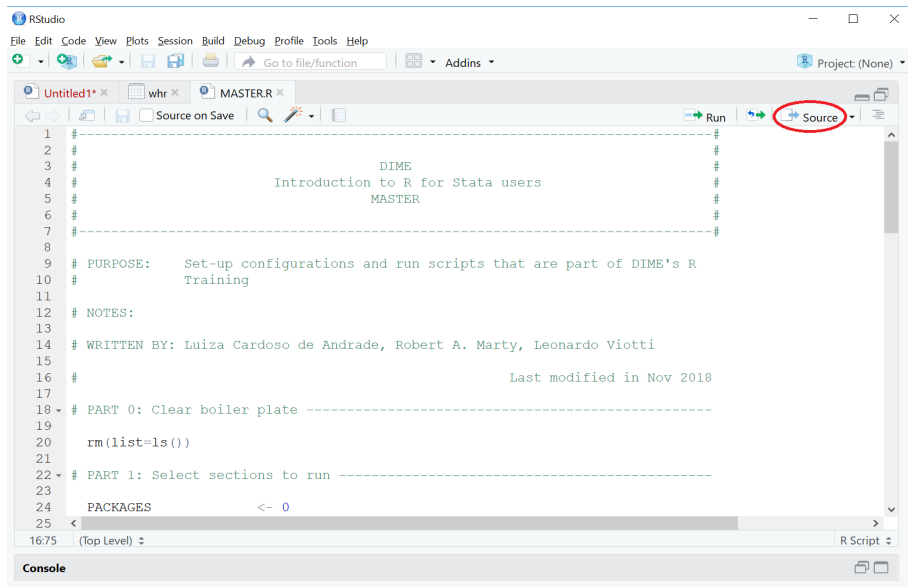
RStudio interface



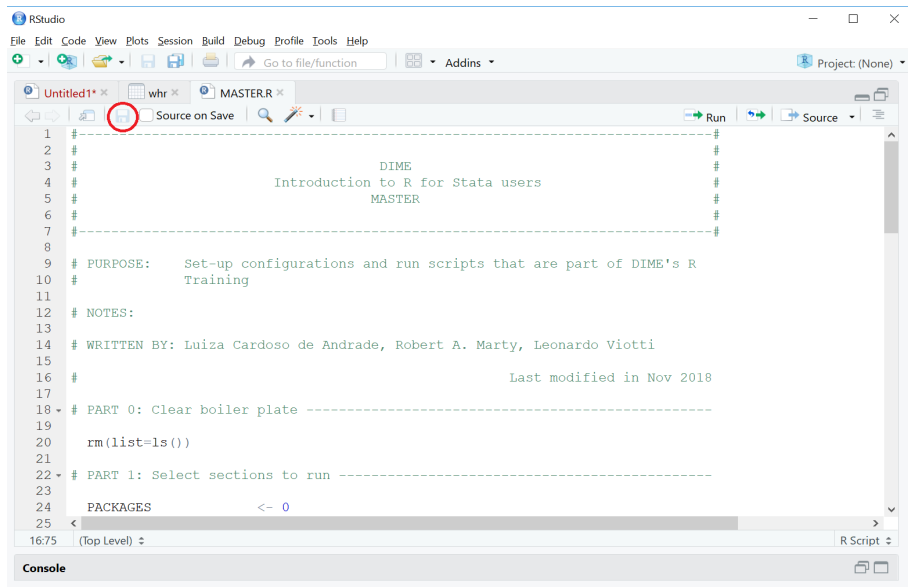
RStudio interface



RStudio interface



RStudio interface



Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R**
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

- In Stata, you can open ONE dataset, and perform operations that can change this dataset.
- You can also have other objects, such as matrices, macros and tempfiles, but they are secondary, and most functions only use the main dataset.
- If you wish to do any non-permanent changes to your data, you'll need to preserve the original data to keep it intact.
- R works in a completely different way: you can have as many datasets (objects) as you wish (or your computer's memory allows) and operations will only have lasting effects if you store them.

- Everything that exists in R's memory – variables, datasets, functions – is an object.
- An object is a chunk of data stored in the memory that has a name by which you call it (exactly like macros in Stata).
- If you create an object, it is going to be stored in memory until you delete it or quit R.
- Whenever you run anything you intend to use in the future, you need to store it as an object.

To better understand the idea, we're going to use the data from the United Nations' World Happiness Report. First, let's take a look at the data.

Type the following code to explore the data:

```
# We can use the function View() to browse the whole data  
View(whr)  
  
# Alternatively we can print the first 6 obs. with head()  
head(whr)
```

Data in R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

Untitled1* whr MASTER.R

Filter

	country	region	year	happy_rank	happy_score	gdp_pc	family	health	freedom
1	Norway	Western Europe	2017	1	7.537	1.61646318	1.5335236	0.796666503	0.6354225
2	Denmark	Western Europe	2017	2	7.522	1.48238301	1.5511216	0.792565525	0.6260067
3	Iceland	Western Europe	2017	3	7.504	1.48063302	1.6105740	0.833552122	0.6271626
4	Switzerland	Western Europe	2017	4	7.494	1.56497955	1.5169117	0.858131289	0.6200705
5	Finland	Western Europe	2017	5	7.469	1.44357193	1.5402467	0.809157670	0.6179508
6	Netherlands	Western Europe	2017	6	7.377	1.50394464	1.4289392	0.810696125	0.5853844
7	Canada	North America	2017	7	7.316	1.47920442	1.4813490	0.834557652	0.6111009
8	New Zealand	Australia and New Zealand	2017	8	7.314	1.40570605	1.5481951	0.816759706	0.6140621
9	Sweden	Western Europe	2017	9	7.284	1.49438727	1.4781622	0.830875158	0.6129241
10	Australia	Australia and New Zealand	2017	10	7.284	1.48441494	1.5100420	0.843886793	0.6016073
11	Israel	Middle East and Northern Africa	2017	11	7.213	1.37538242	1.3762900	0.838404000	0.4059886
12	Costa Rica	Latin America and Caribbean	2017	12	7.079	1.10970628	1.4164037	0.759509265	0.5801316
13	Austria	Western Europe	2017	13	7.006	1.48709726	1.4599450	0.815328419	0.5677661
14	United States	North America	2017	14	6.993	1.54625928	1.4199206	0.774286628	0.5057405
15	Ireland	Western Europe	2017	15	6.977	1.52570664	1.5582211	0.809787634	0.5721102

Showing 1 to 16 of 469 entries

Console

Data in R

```
head(whr)
```

```
##      country      region year happy_rank happy_score  gdp_pc  family
## 1      Norway Western Europe 2017         1      7.537 1.616463 1.533524
## 2      Denmark Western Europe 2017         2      7.522 1.482383 1.551122
## 3      Iceland Western Europe 2017         3      7.504 1.480633 1.610574
## 4 Switzerland Western Europe 2017         4      7.494 1.564980 1.516912
## 5      Finland Western Europe 2017         5      7.469 1.443572 1.540247
## 6 Netherlands Western Europe 2017         6      7.377 1.503945 1.428939
##      health  freedom trust_gov_corr generosity dystopia_res
## 1 0.7966665 0.6354226      0.3159638 0.3620122      2.277027
## 2 0.7925655 0.6260067      0.4007701 0.3552805      2.313707
## 3 0.8335521 0.6271626      0.1535266 0.4755402      2.322715
## 4 0.8581313 0.6200706      0.3670073 0.2905493      2.276716
## 5 0.8091577 0.6179509      0.3826115 0.2454828      2.430182
## 6 0.8106961 0.5853845      0.2826618 0.4704898      2.294804
```

Now, let's try some simple manipulations. First, assume we're only interested in data of the year 2015.

Exercise 2: Subset the data

- 1 Subset the data set, keeping only observations where variable year equals 2015.

To do that we'll use the subset() function
`subset(whr, year == 2015)`

- 2 Then, look again at the first 6 observations

Use the head() function again
`head(whr)`

Data in R

```
head(whr)
```

```
##      country      region year happy_rank happy_score  gdp_pc  family
## 1      Norway Western Europe 2017         1      7.537 1.616463 1.533524
## 2      Denmark Western Europe 2017         2      7.522 1.482383 1.551122
## 3      Iceland Western Europe 2017         3      7.504 1.480633 1.610574
## 4 Switzerland Western Europe 2017         4      7.494 1.564980 1.516912
## 5      Finland Western Europe 2017         5      7.469 1.443572 1.540247
## 6 Netherlands Western Europe 2017         6      7.377 1.503945 1.428939
##      health  freedom trust_gov_corr generosity dystopia_res
## 1 0.7966665 0.6354226      0.3159638 0.3620122      2.277027
## 2 0.7925655 0.6260067      0.4007701 0.3552805      2.313707
## 3 0.8335521 0.6271626      0.1535266 0.4755402      2.322715
## 4 0.8581313 0.6200706      0.3670073 0.2905493      2.276716
## 5 0.8091577 0.6179509      0.3826115 0.2454828      2.430182
## 6 0.8106961 0.5853845      0.2826618 0.4704898      2.294804
```

We can see that nothing happened to the original data. This happens because we didn't store the edit we made anywhere.

To store an object, we use the assignment operator (`<-`):

```
# Assign the Answer to the Ultimate Question of Life,  
# the Universe, and Everything  
x <- 42
```

From now on, `x` is associated with the stored value (until you replace it delete it or close R).

Exercise 3: Create an object

Create a new data set, called `whr2015`, that is a subset of the `whr` data set containing only data from the year 2015.

Using the same function but now assigning it to an object

```
whr2015 <- subset(whr, year == 2015)
```

Display the 5 first obs. of the new data

```
head(whr2015)
```

Notice that we still have the original data set intact

```
head(whr)
```

Data in R

```
head(whr2015)
```

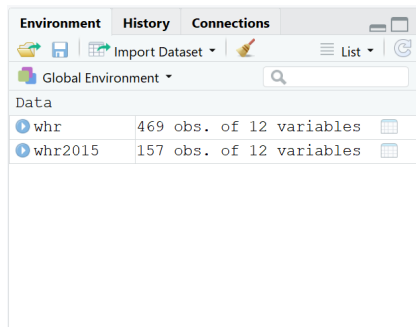
```
##      country      region year happy_rank happy_score  gdp_pc  family
## 313 Switzerland Western Europe 2015         1      7.587 1.39651 1.34951
## 314    Iceland Western Europe 2015         2      7.561 1.30232 1.40223
## 315    Denmark Western Europe 2015         3      7.527 1.32548 1.36058
## 316     Norway Western Europe 2015         4      7.522 1.45900 1.33095
## 317     Canada North America 2015         5      7.427 1.32629 1.32261
## 318     Finland Western Europe 2015         6      7.406 1.29025 1.31826
##      health freedom trust_gov_corr generosity dystopia_res
## 313 0.94143 0.66557      0.41978      0.29678      2.51738
## 314 0.94784 0.62877      0.14145      0.43630      2.70201
## 315 0.87464 0.64938      0.48357      0.34139      2.49204
## 316 0.88521 0.66973      0.36503      0.34699      2.46531
## 317 0.90563 0.63297      0.32957      0.45811      2.45176
## 318 0.88911 0.64169      0.41372      0.23351      2.61955
```


Data in R

```
head(whr)
```

```
##      country      region year happy_rank happy_score  gdp_pc  family
## 1      Norway Western Europe 2017         1      7.537 1.616463 1.533524
## 2      Denmark Western Europe 2017         2      7.522 1.482383 1.551122
## 3      Iceland Western Europe 2017         3      7.504 1.480633 1.610574
## 4 Switzerland Western Europe 2017         4      7.494 1.564980 1.516912
## 5      Finland Western Europe 2017         5      7.469 1.443572 1.540247
## 6 Netherlands Western Europe 2017         6      7.377 1.503945 1.428939
##      health  freedom trust_gov_corr generosity dystopia_res
## 1 0.7966665 0.6354226      0.3159638 0.3620122      2.277027
## 2 0.7925655 0.6260067      0.4007701 0.3552805      2.313707
## 3 0.8335521 0.6271626      0.1535266 0.4755402      2.322715
## 4 0.8581313 0.6200706      0.3670073 0.2905493      2.276716
## 5 0.8091577 0.6179509      0.3826115 0.2454828      2.430182
## 6 0.8106961 0.5853845      0.2826618 0.4704898      2.294804
```

You can also see that your environment pane now has two objects:



Two important concepts to take note:

- ❶ In R, if you want to change your data, you need to store it in an object.
- ❷ It is possible to simply replace the original data. This happens if you assign the new object to the same name as the original.
- ❸ Print (display) is built into R. If you execute any action without storing it, R will simply print the results of that action but won't save anything in the memory.

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects**
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

Objects are the building blocks of R programming. This section will explore some of the most common.

This will give you the foundation to explore your data and construct analytical outputs.

Here are the types of object we will cover today:

- **Vectors:** an uni-dimensional object that stores a sequence of values
- **Data frames:** a combination of different vectors of the same length (the same as your data set in Stata)
- **Lists:** a multidimensional object that can store several objects of different dimension

R objects

Vectors

A vector is an uni-dimensional object composed by one or more scalars of the same type.

Use the following code to create vectors in two different ways

```
# Creating a vector with the c() function
```

```
v1 <- c(1,1,2,3,5)
```

```
# Alternative way to create an evenly spaced vector
```

```
v2 <- 1:5
```

You can use brackets for indexing

```
# Print the 4th element of the vector
```

```
v2[4]
```

```
## [1] 4
```

R objects

Vectors

To R, each of the columns of `whr` is a vector.

Calling a vector from a `data.frame` column

We use the `$` to call vectors (variables) by their names in a `data.frame`

Type the following code:

```
# Create a vector with the values of the `year` variable  
year_vec <- whr$year  
  
# See the 3 first elements of the year column  
whr$year[1:3]  
  
## [1] 2017 2017 2017
```


R objects

Data Frames

The `whr` and `whr2015` objects are both data frames. You can also construct a new data frame from scratch by combining vectors with the same number of elements .

Now, type the following code to create a new data frame

```
# Dataframe created by binding vectors
```

```
df1 <- data.frame(v1,v2)
```

```
df1
```

```
##      v1 v2
## 1     1  1
## 2     1  2
## 3     2  3
## 4     3  4
## 5     5  5
```

Since a data frame has two dimensions, you can use indexing on both:

Numeric indexing

The first column of whr

```
whr[,1]
```

The 45th line of whr

```
whr[45,]
```

Or the 45th element of the first line

```
whr[45,1]
```

Alternatively, you can use the column names for indexing, which is the same as using the \$ sign.

Names indexing

```
# Or the 22th element of the country column  
whr[22, "country"] # The same as whr$country[22]
```

```
## [1] Brazil
```

```
## 164 Levels: Afghanistan Albania Algeria Angola Argentina ..
```

Lists are more complex objects that can contain many objects of different classes and dimensions.

Lists are fancy and can have a lot of functionalities and attributes. They are the output of many functions and are used to construct complex objects.

It would be beyond the scope of this introduction to go deep into them, but here's a quick example

Combine several objects of different types in a list

```
# Use the list() function  
lst <- list(v1, df1, 45)
```

Print the list yourself to see how it looks like.

R objects

Lists

```
# Check the contents of lst
```

```
print(lst)
```

```
## [[1]]
```

```
## [1] 1 1 2 3 5
```

```
##
```

```
## [[2]]
```

```
##      v1 v2
```

```
## 1    1  1
```

```
## 2    1  2
```

```
## 3    2  3
```

```
## 4    3  4
```

```
## 5    5  5
```

```
##
```

```
## [[3]]
```

```
## [1] 45
```

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data**
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

Basic types of data

R has different kinds of data that can be recorded inside objects. They are very similar to what you have in Stata, and the main types are string, integer, numeric, factor and boolean.

Let's start with the simpler ones:

Strings

A sequence of characters and are usually represented between double quotes. They can contain single letters, words, phrases or even some longer text.

Integer and numeric

As in Stata, these are two different ways to store numbers. They are different because they use memory differently. As default, R stores numbers in the numeric format (double).

Basic types of data

Strings

Now we'll use string data to practice some basic object manipulations in R.

Exercise 4: Create a vector of strings

Create two string vector containing the names of commonly used statistical software in order of importance:

```
# Creating string vector  
str_vec <- c("R",  
             "Python",  
             "SAS",  
             "Excel",  
             "Stata")
```

Now print them to check them out.

Basic types of data

Strings

Exercise 5: Concatenate strings

- 1 Create a scalar (a vector of one element) containing the phrase “is better than” and call it `str_scalar`.
- 2 Use the function `paste()` with 3 arguments separated by commas:
 - The first argument as the 1st element of `str_vec`.
 - The second argument as the `str_scalar`.
 - The third argument as the 5th element of `str_vec`.
- 3 If you're not sure where to start, type:

```
help(paste)
```

Basic types of data

Strings

```
### Using the paste function to combine strings
```

```
# Scalar
```

```
str_scalar <- "is better than"
```

```
# Using the paste() function
```

```
paste(str_vec[1], str_scalar, str_vec[5])
```

```
## [1] "R is better than Stata"
```

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data**
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix

Advanced types of data

R also has other more complex ways of storing data. These are the most used:

Factors

Factors are numeric categorical values with text label, equivalent to labelled variables in Stata. Turning strings into factors makes it easier to run different analyses on them and also uses less space in your memory.

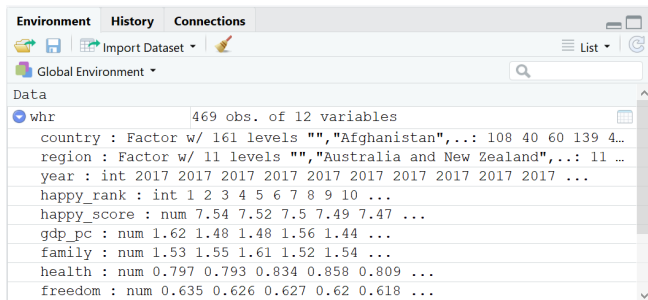
Booleans

Booleans are logical binary variables, accepting either `TRUE` or `FALSE` as values. They are automatically generated when performing logical operations

Advanced types of data

Factors

In `whr`, we can see that `country` and `region` are factor variables. In your environment panel, there is the information of the type of all variables, and for factors the number of levels.



The screenshot shows the RStudio Environment panel with the 'Global Environment' selected. The 'Data' section lists the 'whr' dataset, which contains 469 observations of 12 variables. The variables and their types are as follows:

Variable	Type	Levels / Values
country	Factor	w/ 161 levels: "", "Afghanistan", ...
region	Factor	w/ 11 levels: "", "Australia and New Zealand", ...
year	int	2017
happy_rank	int	1 2 3 4 5 6 7 8 9 10 ...
happy_score	num	7.54 7.52 7.5 7.49 7.47 ...
gdp_pc	num	1.62 1.48 1.48 1.56 1.44 ...
family	num	1.53 1.55 1.61 1.52 1.54 ...
health	num	0.797 0.793 0.834 0.858 0.809 ...
freedom	num	0.635 0.626 0.627 0.62 0.618 ...

Advanced types of data

Factors

We'll learn how to deal with factors in detail on the next session, since they are very important for us. For now, here are two important things to keep in mind when using them:

Warning:

Unlike Stata, in R

- 1 You use the labels to refer to factors
- 2 You cannot choose the underlying values

Advanced types of data

Booleans

Boolean data is the result of logical conditions

- Whenever you're using an `if` statement in Stata, you're implicitly using boolean data.
- The difference is that in R, this can be done in 2 steps.

Advanced types of data

Booleans

Exercise 6:

Create boolean vector with the condition of annual income below average:

```
# Create vector
bool_vec <- (whr$happy_rank <
             mean(whr$happy_rank))

# See the 5 first elements of the vector
head(bool_vec)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```


Advanced types of data

Booleans

Let's use the boolean vector created to add a dummy variable in the `whr` data set for the same condition.

Exercise 6:

- 1 Create a column in `whr` containing zeros and call it `rank_low`. You can do this by typing:

```
whr$rank_low <- 0
```

- 2 Use `bool_vec` to index the lines of the `income_low` column and replace all observations that meet the condition with the value 1.

```
whr$rank_low[bool_vec] <- 1
```

Advanced types of data

Booleans

```
# Replace with 1 those obs that meet the condition  
whr$rank_low [bool_vec] <- 1  
# is the same as  
whr$rank_low[whr$happy_rank < mean(whr$happy_rank)] <- 1  
# This in stata would be  
# replace rank_low = 1 if (...)
```

```
# We can use the summary function to get descriptives  
summary(whr$rank_low)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
## 0.0000  0.0000  0.0000  0.4979  1.0000  1.0000
```

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow**
- 9 Useful resources
- 10 Appendix

Help in R works very much like in Stata: the help files usually start with a brief description of the function, explain its syntax and arguments and list a few examples. There are two ways to access help files:

Exercise 7: Use help

```
# The help() function
```

```
help(summary)
```

```
# and its abbreviation
```

```
?summary
```

- The biggest difference, however, is that R has a much wider user community and it has a lot more online resources.
- For instance, in 2014, Stata had 11 dedicated blogs written by users, while R had 550.¹
- The most powerful problem-solving tool in R, however, is Google. Searching the something yields tons of results.
- Often that means a Stack Overflow page where someone asked the same question and several people gave different answers. Here's a typical example: <https://stackoverflow.com/questions/1660124/how-to-sum-a-variable-by-group>

¹Check <http://r4stats.com/articles/popularity/> for more.

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources**
- 10 Appendix

Useful resources

Blogs and online courses:

- Surviving graduate econometrics with R:
<https://thetarzan.wordpress.com/2011/05/24/surviving-graduate-econometrics-with-r-the-basics-1-of-8/>
- An Introduction to R at <https://cran.r-project.org/>
- R programming in Coursera:
<https://www.coursera.org/learn/r-programming>
- Try R in Code School: <http://tryr.codeschool.com/>
- R programming for dummies:
<http://www.dummies.com/programming/r/>
- R bloggers: <https://www.r-bloggers.com/>
- R statistics blog: <https://www.r-statistics.com/>
- The R graph gallery: <https://www.r-graph-gallery.com/>

Books:

- R for Stata Users - Robert A. Muenchen and Joseph Hilbe
- R Graphics Cookbook - Winston Chang
- R for Data Science - Hadley Wickha and Garrett Grolemond

Thank you!

Outline

- 1 Introduction
- 2 Getting started
- 3 RStudio interface
- 4 Data in R
- 5 R objects
- 6 Basic types of data
- 7 Advanced types of data
- 8 Help, Google and Stackoverflow
- 9 Useful resources
- 10 Appendix**

R's syntax is a bit heavier than Stata's:

- Parentheses to separate function names from its arguments.
- Commas to separate arguments.
- For comments we use the # sign.
- You can have line breaks inside function statements.
- In R, functions can be treated much like any other object Therefore, they can be passed as arguments to other functions.

Similarly to Stata:

- Square brackets are used for indexing.
- Curly braces are used for loops and if statements.
- Largely ignores white spaces.

Appendix - RStudio interface

Script

Where you write your code. Just like a do file.

Console

Where your results and messages will be displayed. But you can also type commands directly into the console, as in Stata.

Environment

What's in R's memory.

The 4th pane

Can display different things, including plots you create, packages loaded and help files.

Appendix - Matrices

A matrix is a bi-dimensional object composed by one or more vectors of the same type.

Type the following code to test two different ways of creating matrices

```
# Matrix created by joining two vectors:
```

```
m1 <- cbind(v1,v1)
```

```
# Matrix using the
```

```
m2 <- matrix(c(1,1,2,3,5,8), ncol = 2)
```

Appendix - Matrices

Now use the following code to check the elements of these matrices by indexing

```
# Matrix indexing: typing matrix[i,j] will give you  
# the element in the ith row and jth column of that matrix  
#m2[1,2]
```

```
# Matrix indexing: typing matrix[i,] will give you the  
# ith row of that matrix  
m1[1,]
```

```
# Matrix indexing: typing matrix[,j] will give you the  
# jth column of that matrix (as a vector)  
m1[,2]
```

Appendix - Advanced types of data - Factors

Factors

Create a factor vector using the following code

```
# Basic factor vector
```

```
num_vec <- c(1,2,2,3,1,2,3,3,1,2,3,3,1)
```

```
fac_vec <- factor(num_vec)
```

```
# A bit fancier factor vector
```

```
fac_vec <- factor(num_vec, labels=c("A", "B", "C"))
```

```
# Change labels
```

```
levels(fac_vec) = c('One', 'Two', 'Three')
```

Appendix - Numbers and integers

Two scalars, one with a round number the other with a fractional part

```
# a numeric scalar with an integer number
```

```
int <- 13
```

```
num <- 12.99
```


Appendix - Numbers and integers

Now we can see the objects classes with the `class()` function and test it with the `is.integer()` and `is.numeric()` functions.

you can see the number's format using the class function:

```
class(int)
```

```
## [1] "numeric"
```

```
class(num)
```

```
## [1] "numeric"
```

you can test the class with the is. method

```
is.integer(int)
```

```
## [1] FALSE
```

Appendix - Numbers and integers

Numbers and integers

We can, however, coerce objects into different classes. We just need to be careful because the result might not be what we're expecting.

Use the *as.integer()* and *round()* functions on the *num* object to see the difference:

```
as.integer(num)
```

```
## [1] 12
```

```
# and
```

```
round(num)
```

```
## [1] 13
```