Introduction to Data Science

Pierre-Alexandre Balland

Key issues

- A data-driven revolution
- The tech giants and the rest of the world (a story of positive feedback loops)
- Key applications of data science (business, policy and science)
- From data to stories
- A lot of data > really precise data
- Algorithmic bias, ethics and echo chambers
- Network Science, Big Data, AI, Machine Learning and Deep Learning

Computer lab: R & RStudio

- In this course we will perform structural network analysis with packages implemented in the R statistical software
- R is the software – but we will use Rstudio as an interface
- R is an open-source project lifted by a virtual community of thousands of developers and million of users worldwide

Why R?

- Reproducibility – R scripts
- Today R offers the most elegant and comprehensive language for the structural and dynamic analysis of networks
- It's free and contains state-of-the-art statistical and graphical routines not yet available in other software
- You can do all your analysis in R, but also data scrapping, create a webpage, or write your research paper

Getting started with R

- Using R is easier than it looks like. And once you master it, you save a ridiculous amount of time

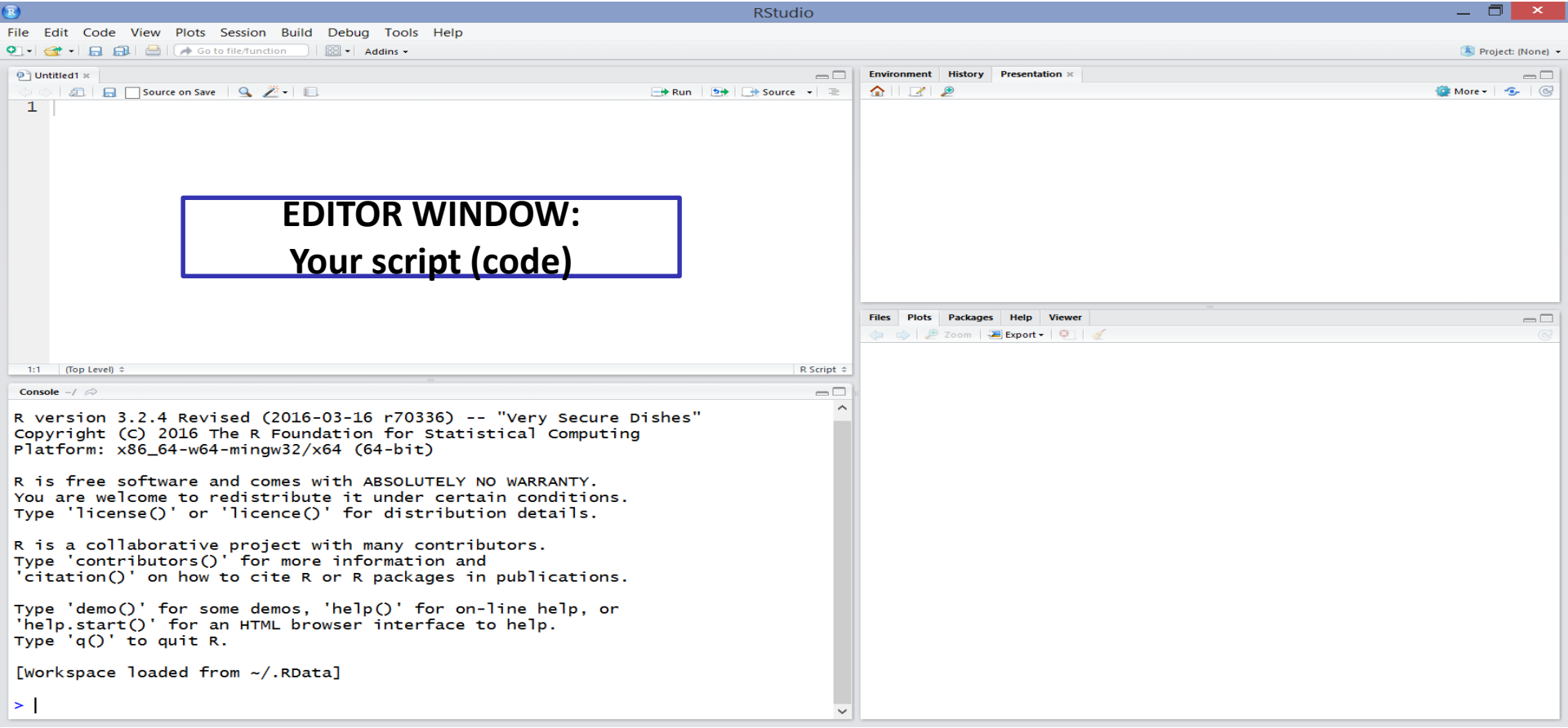
- Afraid of R? It is just a big calculator (a very smart one)

- R is case sensitive

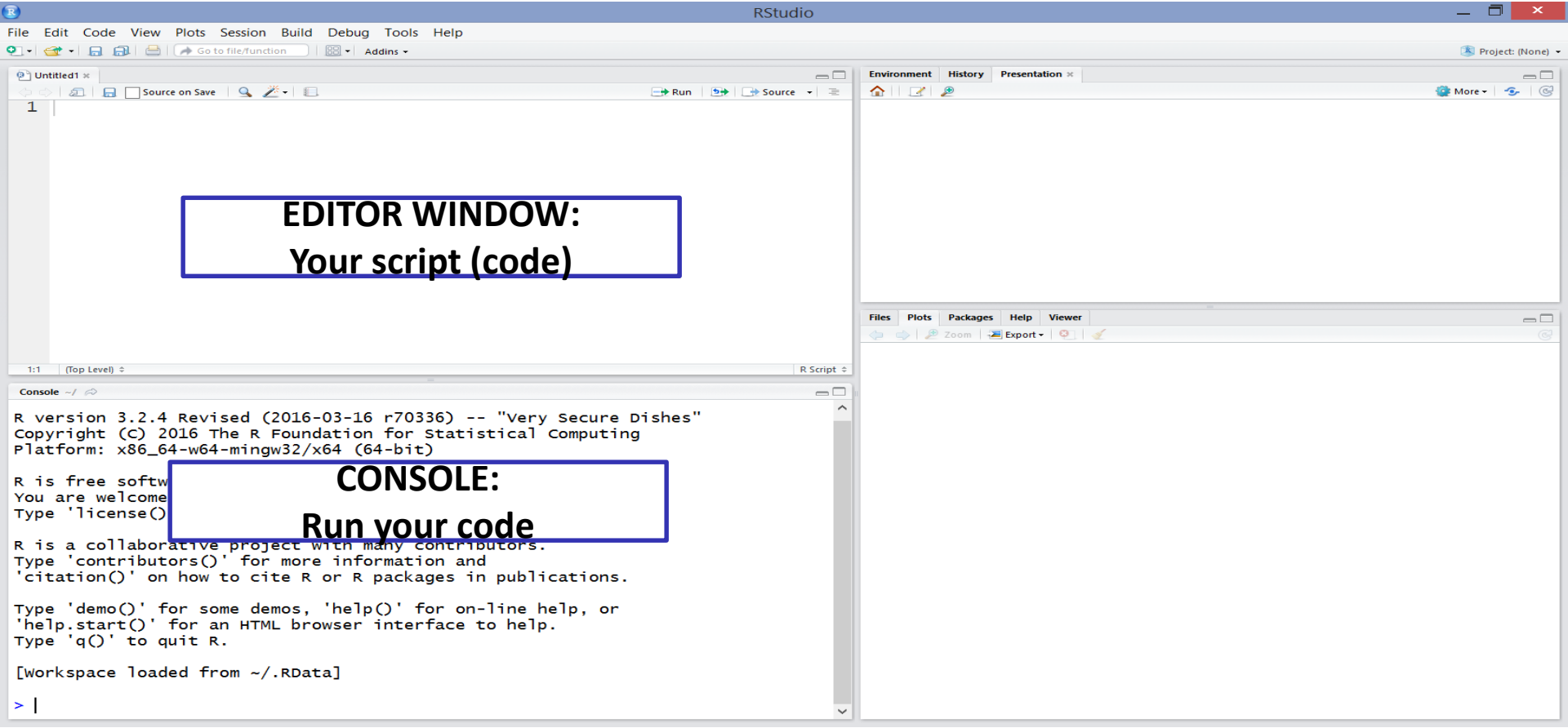
- The # character at the beginning of a line signifies a comment, it is ignored by R

- The <- operator assigns values to objects (data, output of

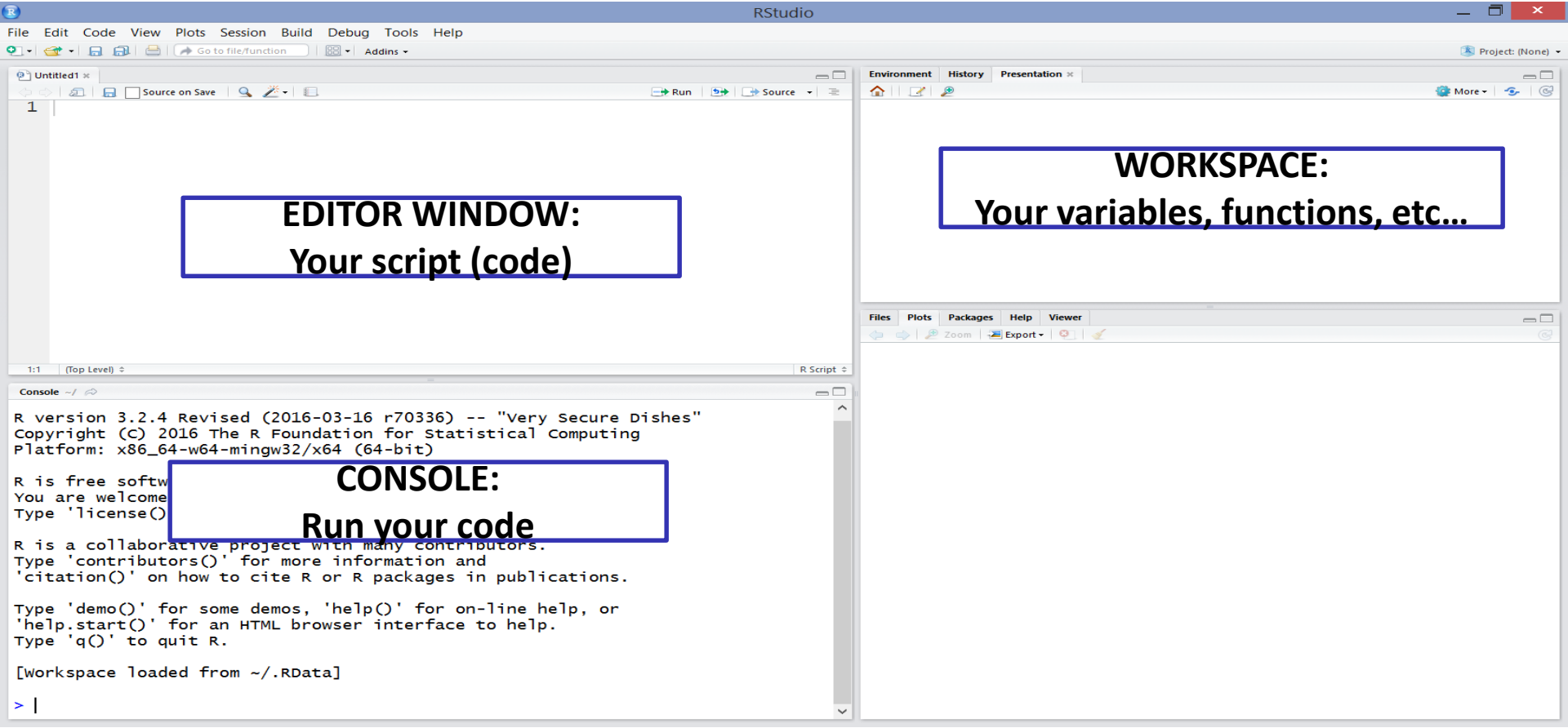
RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
EDITOR WINDOW:
Your script (code)
1:1
(Top Level)
R Script
Console ~/
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment History Presentation
More
Files Plots Packages Help Viewer
Zoom
Export

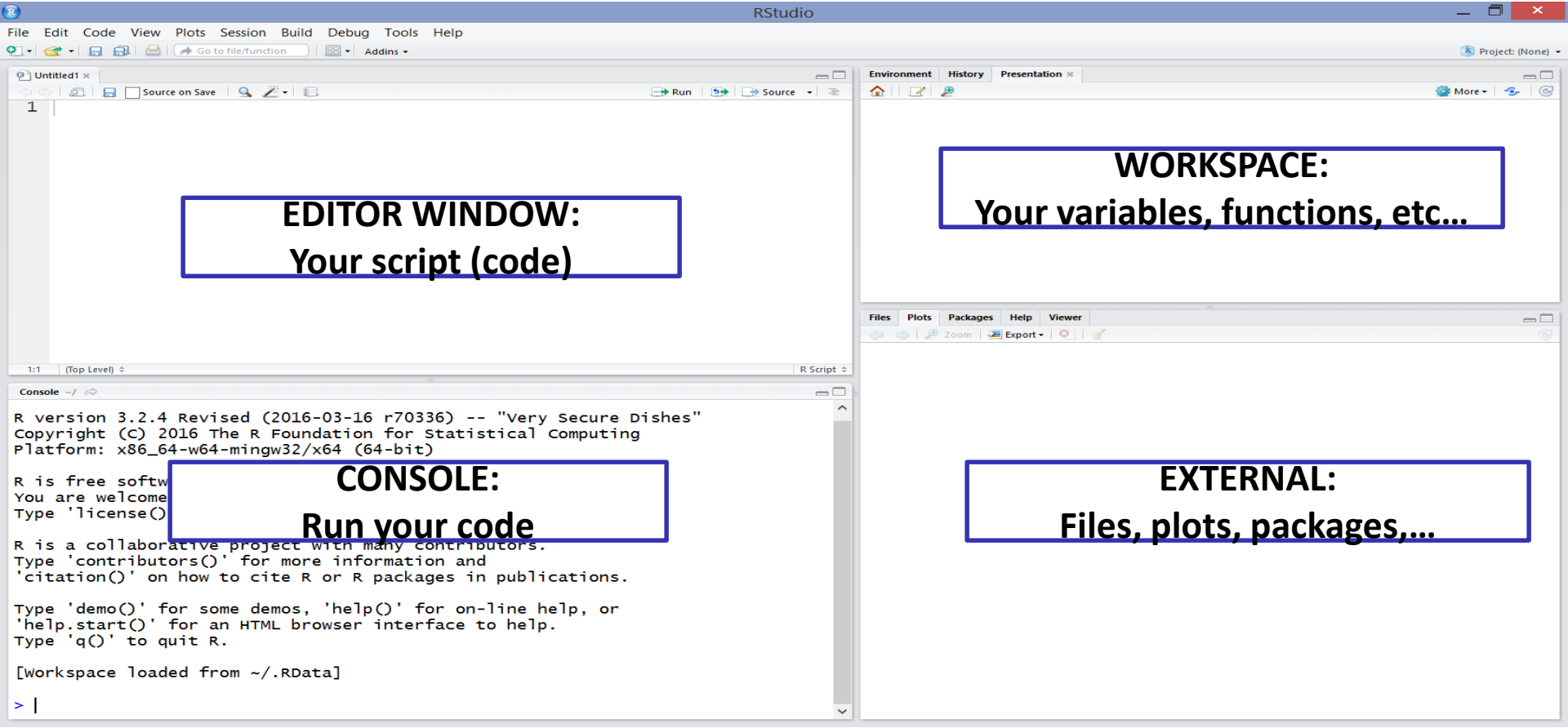
RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
EDITOR WINDOW:
Your script (code)
1:1
(Top Level)
R Script
Console ~/
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free softw
You are welcome
Type 'license()
CONSOLE:
Run your code
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment History Presentation
More
Files Plots Packages Help Viewer
Zoom
Export

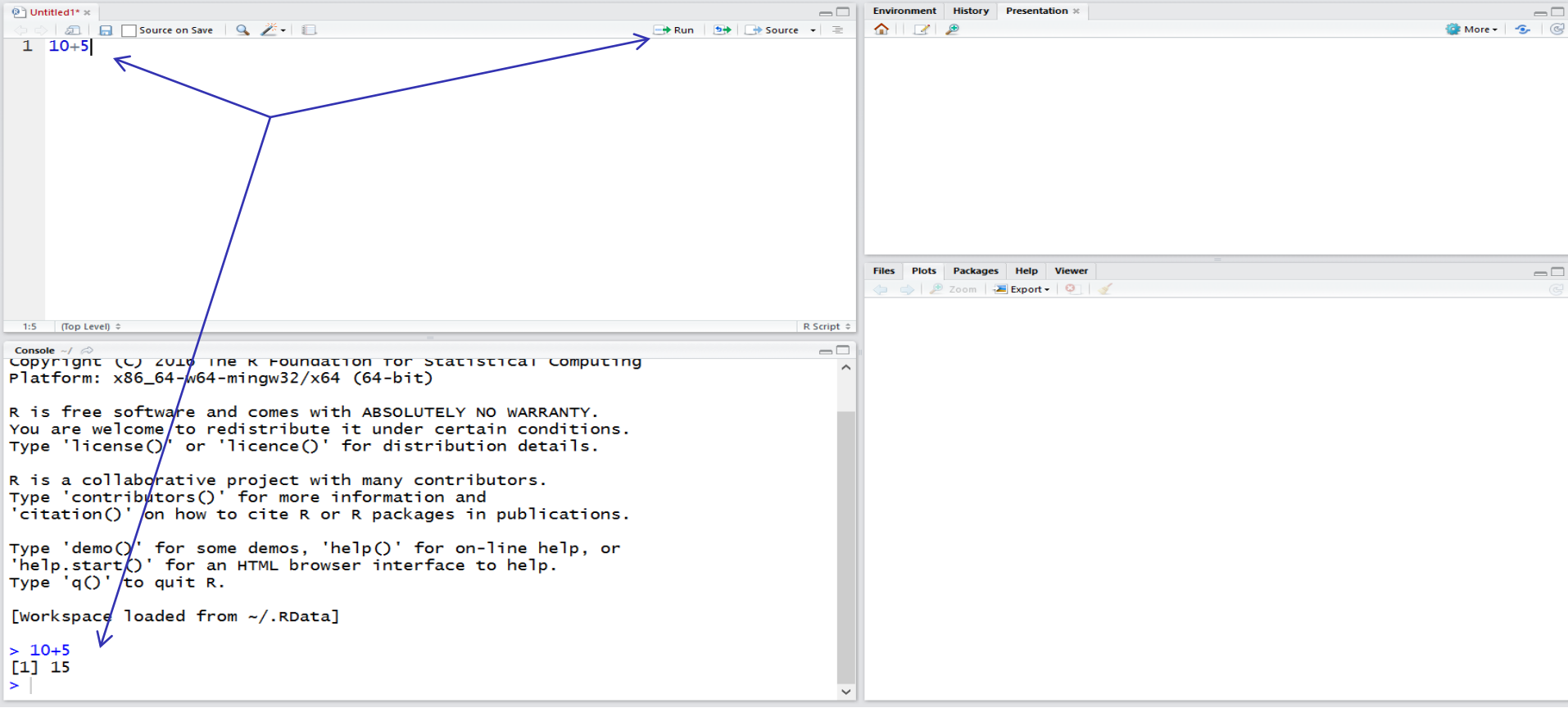
RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
EDITOR WINDOW:
Your script (code)
Environment History Presentation
More
WORKSPACE:
Your variables, functions, etc…
Files Plots Packages Help Viewer
Zoom
Export
Console ~/
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free softw
You are welcome
Type 'license()
CONSOLE:
Run your code
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>

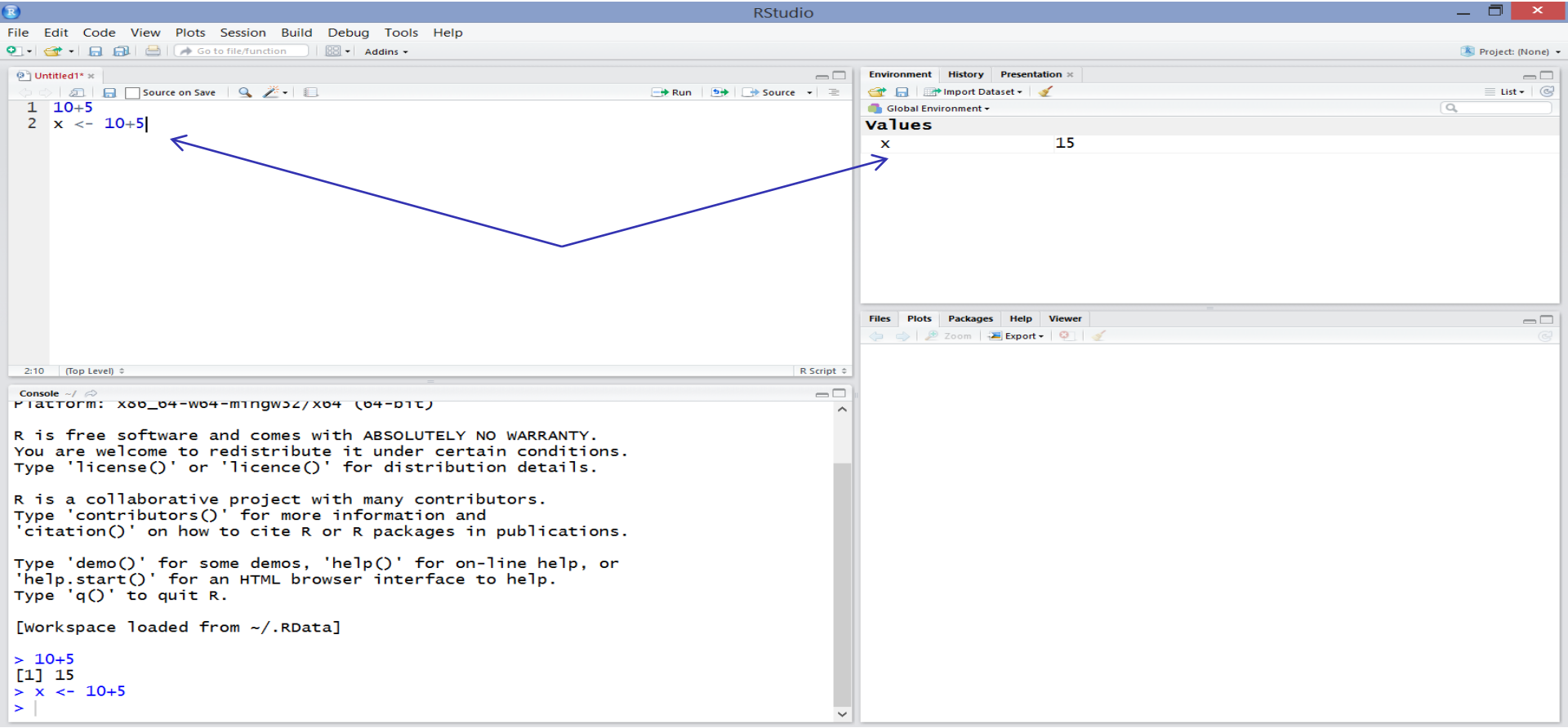
RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
EDITOR WINDOW:
Your script (code)
Environment History Presentation
More
WORKSPACE:
Your variables, functions, etc…
Console ~/
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free softw
You are welcome
Type 'license()
CONSOLE:
Run your code
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Files Plots Packages Help Viewer
Zoom
Export
EXTERNAL:
Files, plots, packages,…

Let’s get started



Create a variable “x”



Create a variable y

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Untitled1*

```
10+5
x <- 10+5
y <- "geography"
```

Console

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> 10+5
[1] 15
> x <- 10+5
> y <- "geography"
>
```

Environment

Values

x	15
y	"geography"

Check the value of x and y

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Untitled1*

```
10+5
x <- 10+5
y <- "geography"
# check the value of x and y
x
y
```

Console

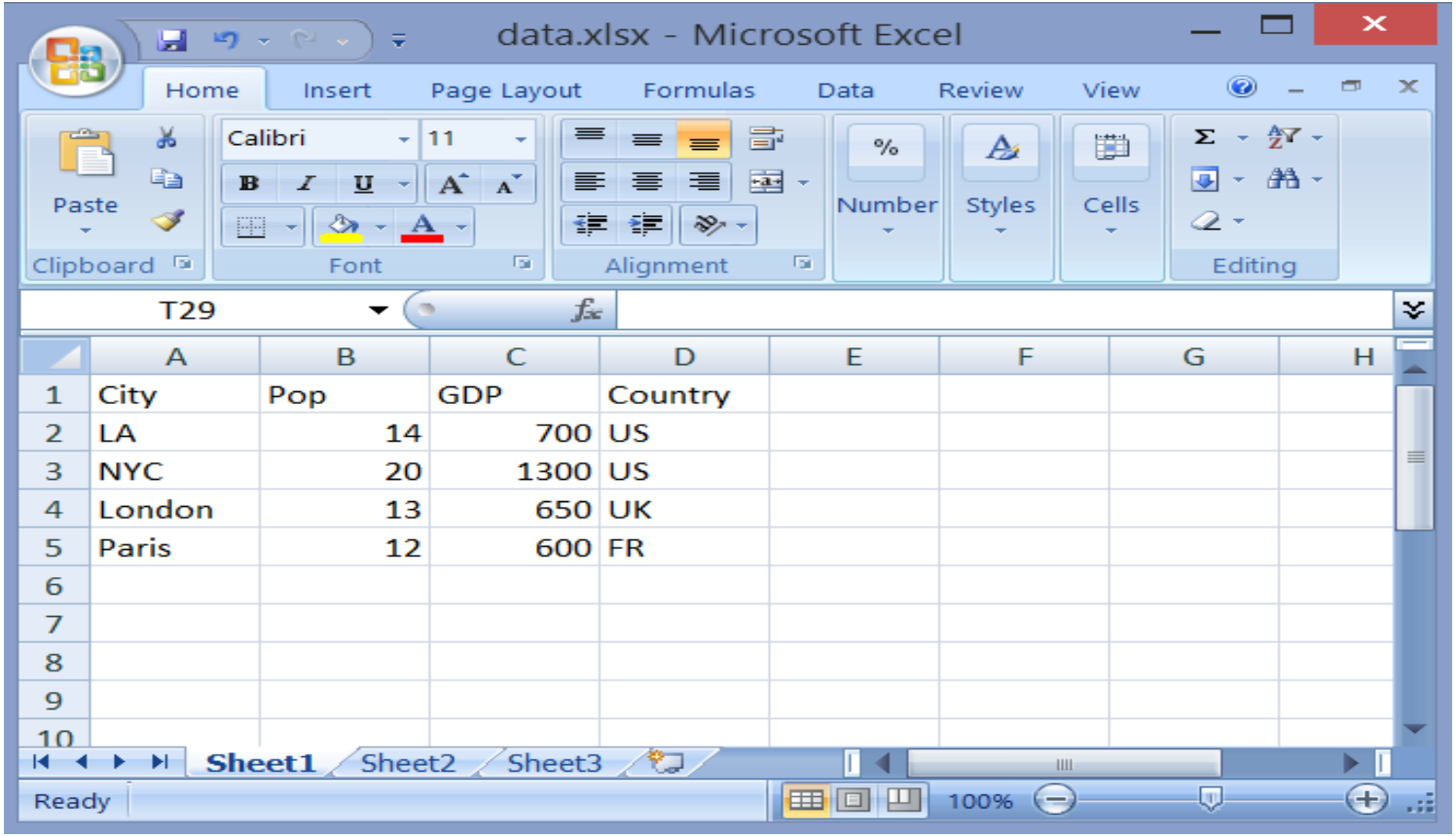
```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> 10+5
[1] 15
> x <- 10+5
> y <- "geography"
> # check the value of x and y
> x
[1] 15
> y
[1] "geography"
>
```

Environment History Presentation

Values

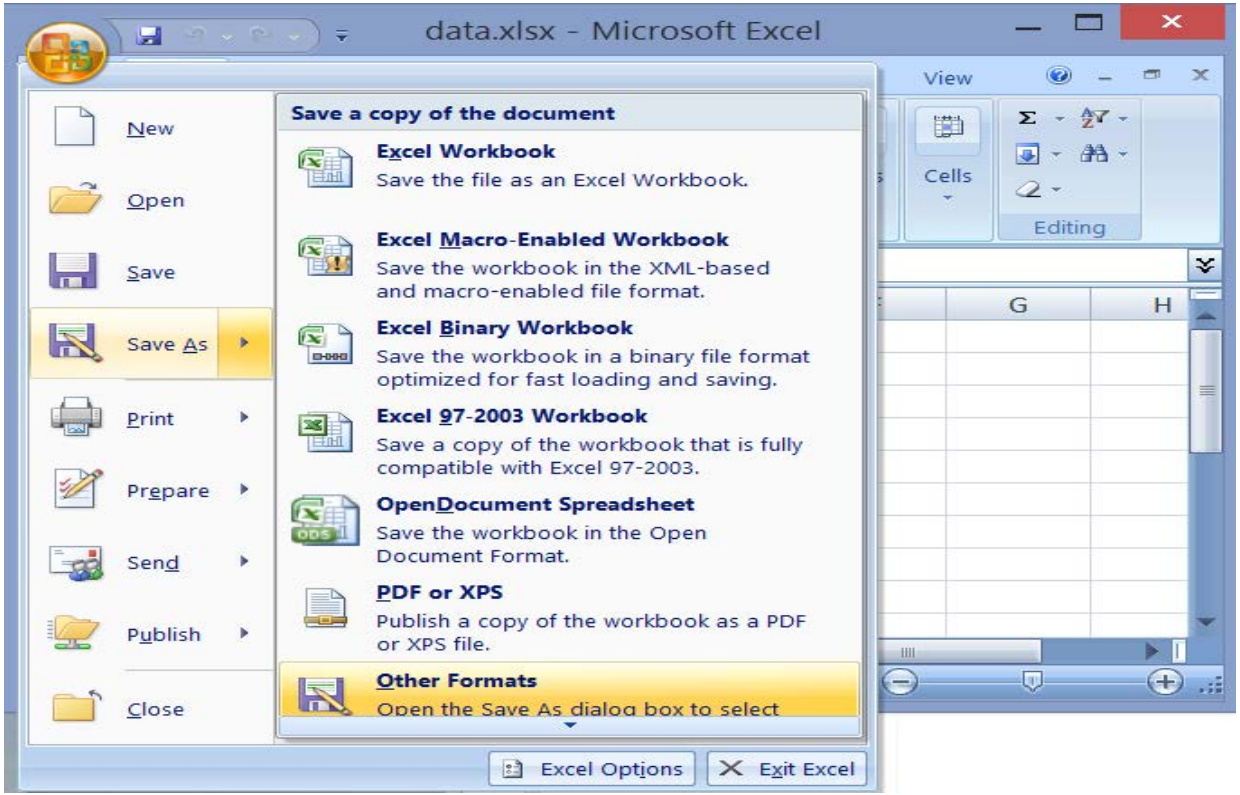
x	15
y	"geography"

Let’s create a toy dataset

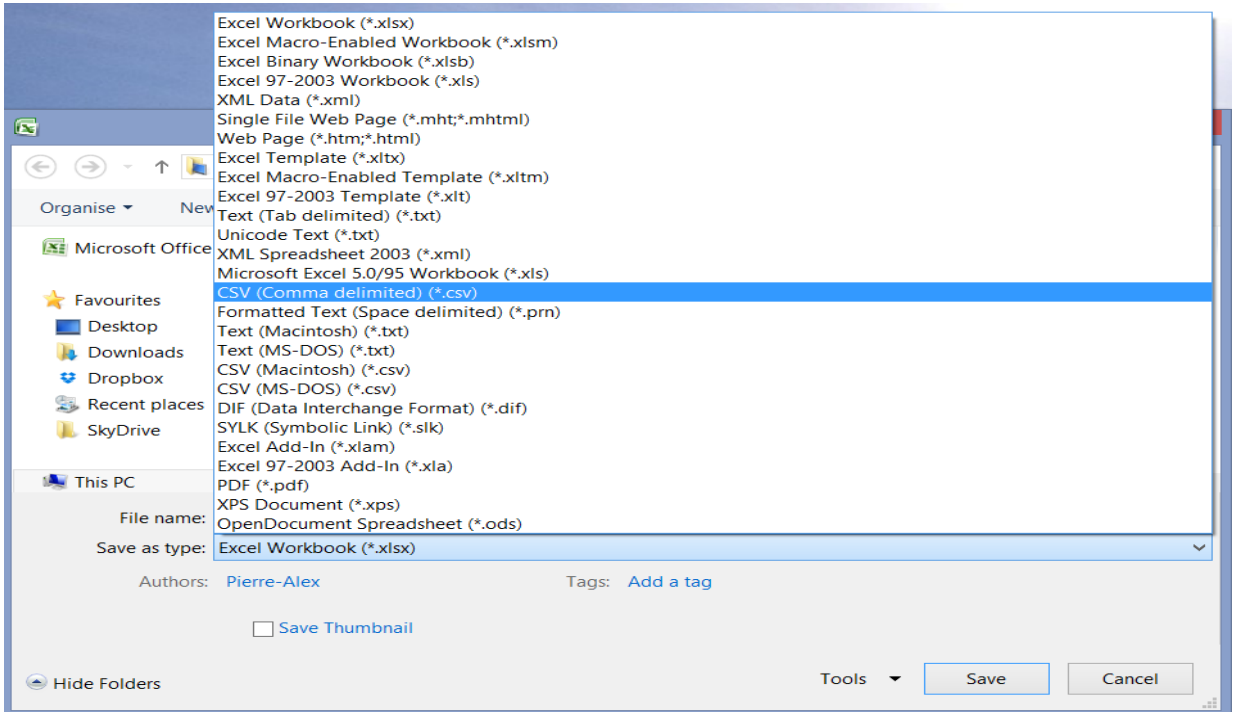


	A	B	C	D
1	City	Pop	GDP	Country
2	LA	14	700	US
3	NYC	20	1300	US
4	London	13	650	UK
5	Paris	12	600	FR

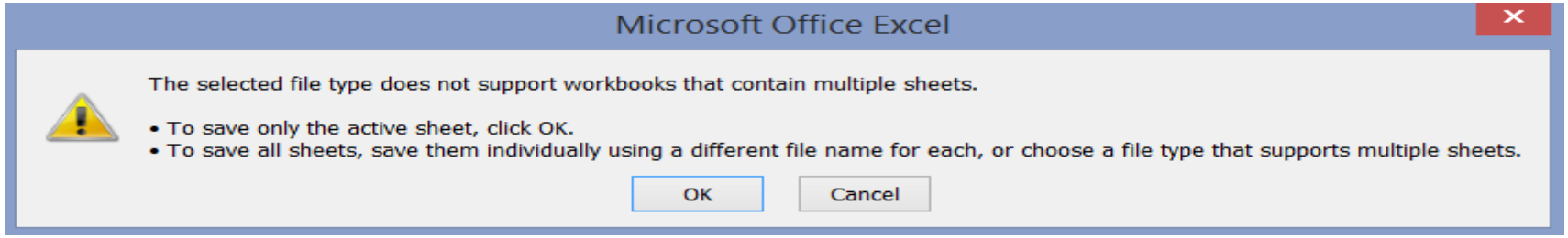
Save as a .csv file



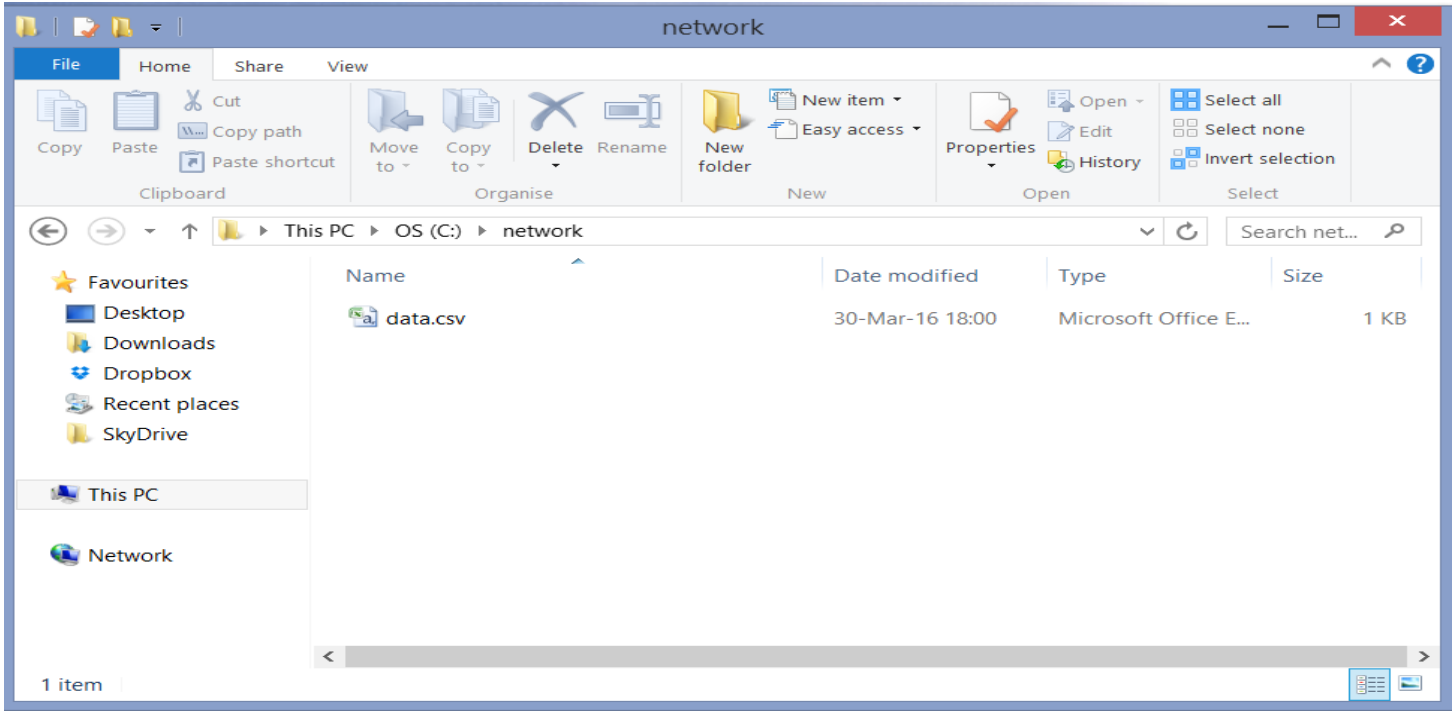
Save as a .csv file



Two warnings: ok



Create a new folder and move the .csv



This is your file path