Exploratory Data Analysis

By
Ravin Poudel
Garrett Lab

Summer 2018
Epidemiology and Data Science

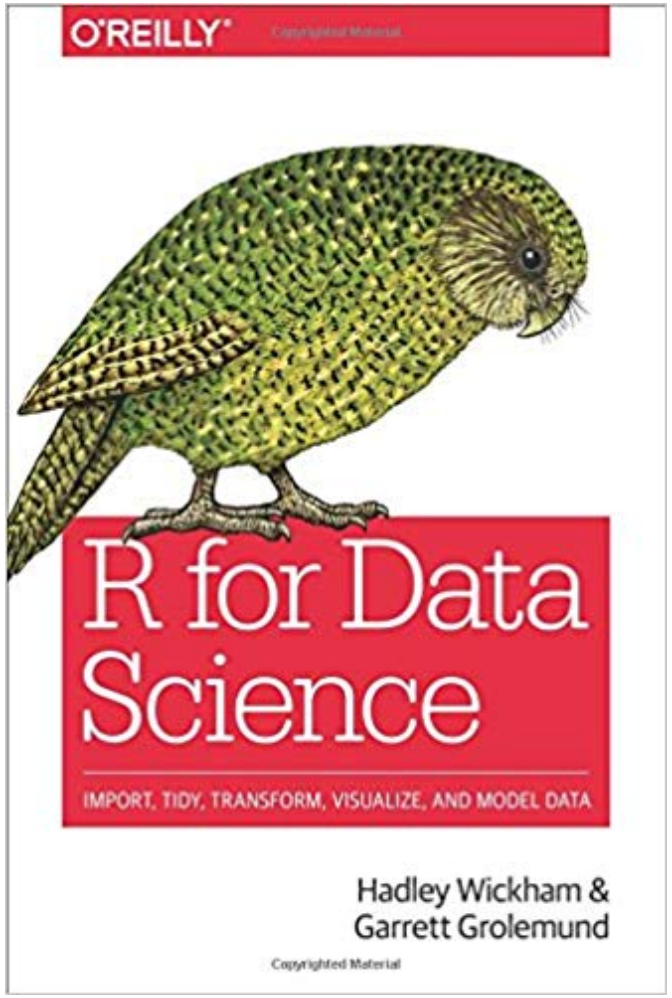
Learning objectives

- Learn to setup R projects and tidy-work environment

- Use visual and data table tools to explore data sets
    - Continuous variable
    - Categorical variables

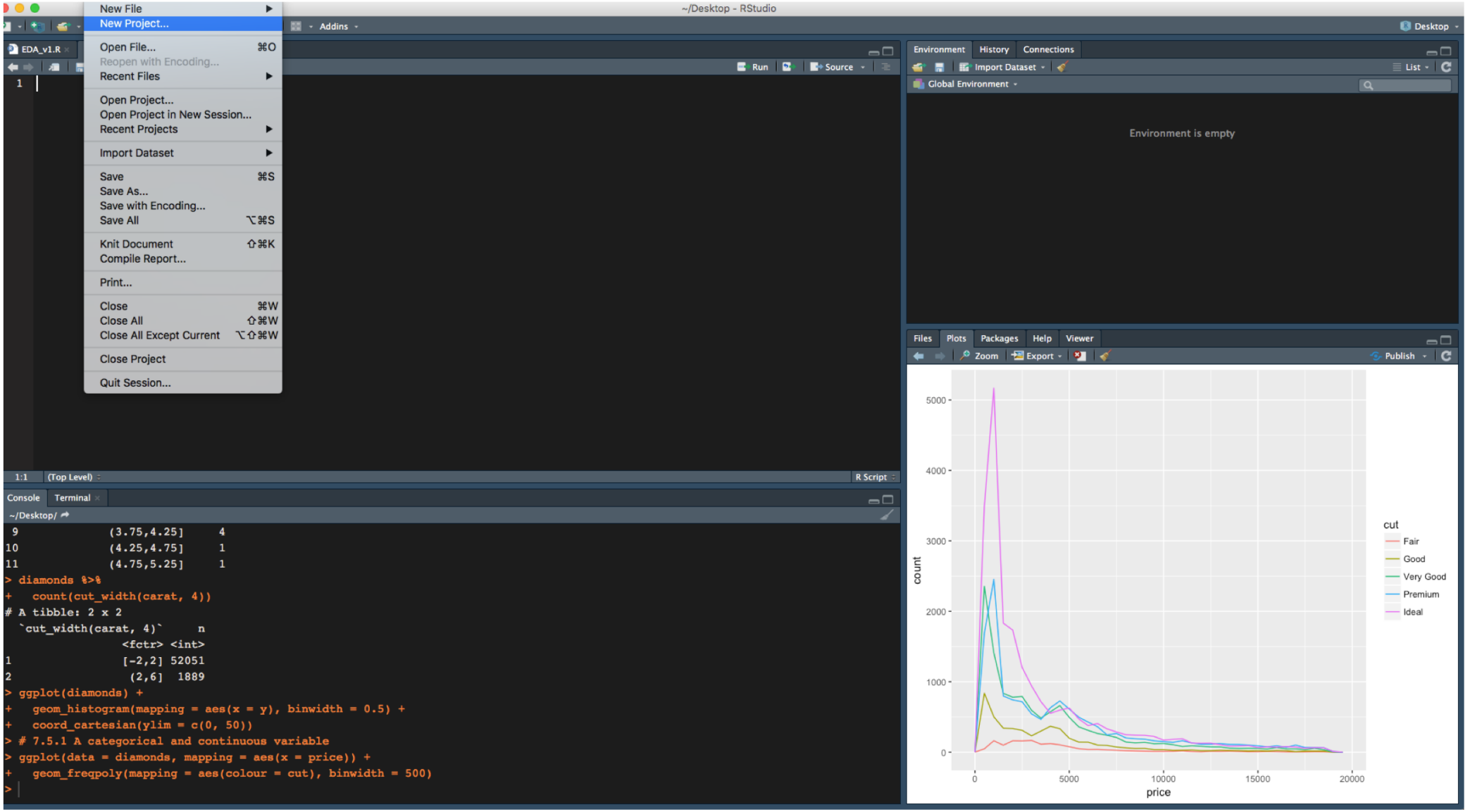
R for Data Science

http://r4ds.had.co.nz/



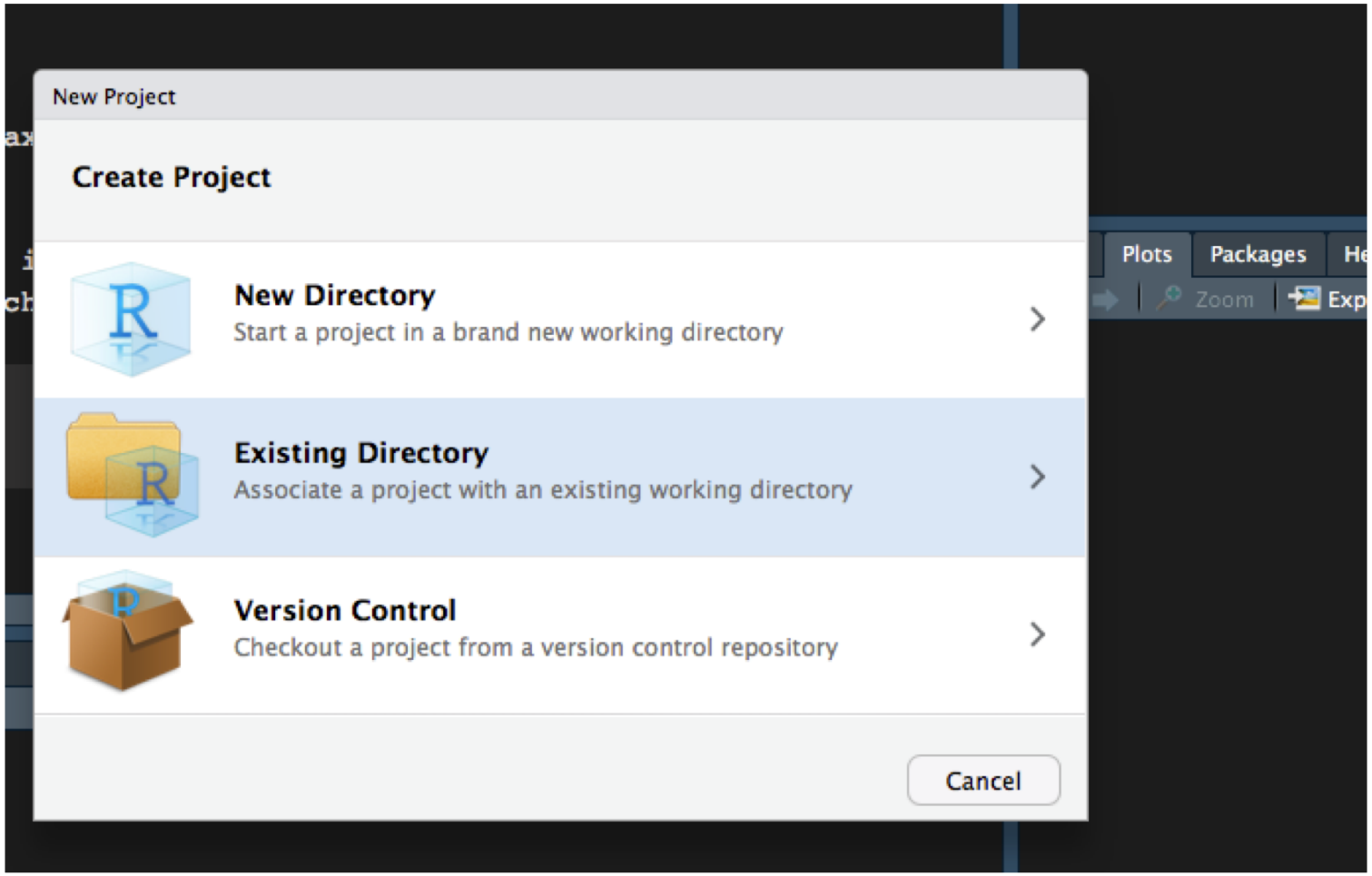
Part One: R project

STEP: 1

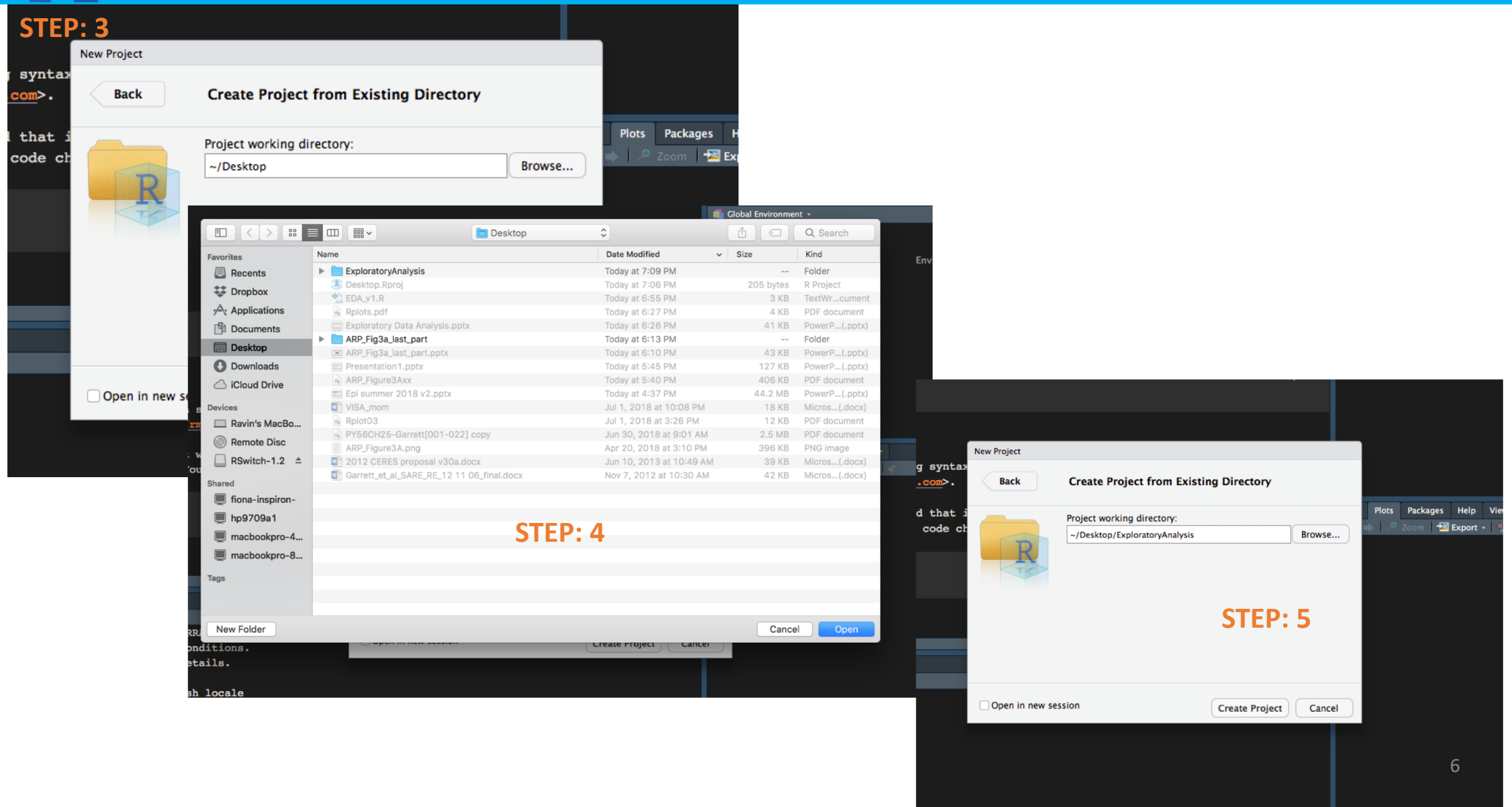


Part One: R project

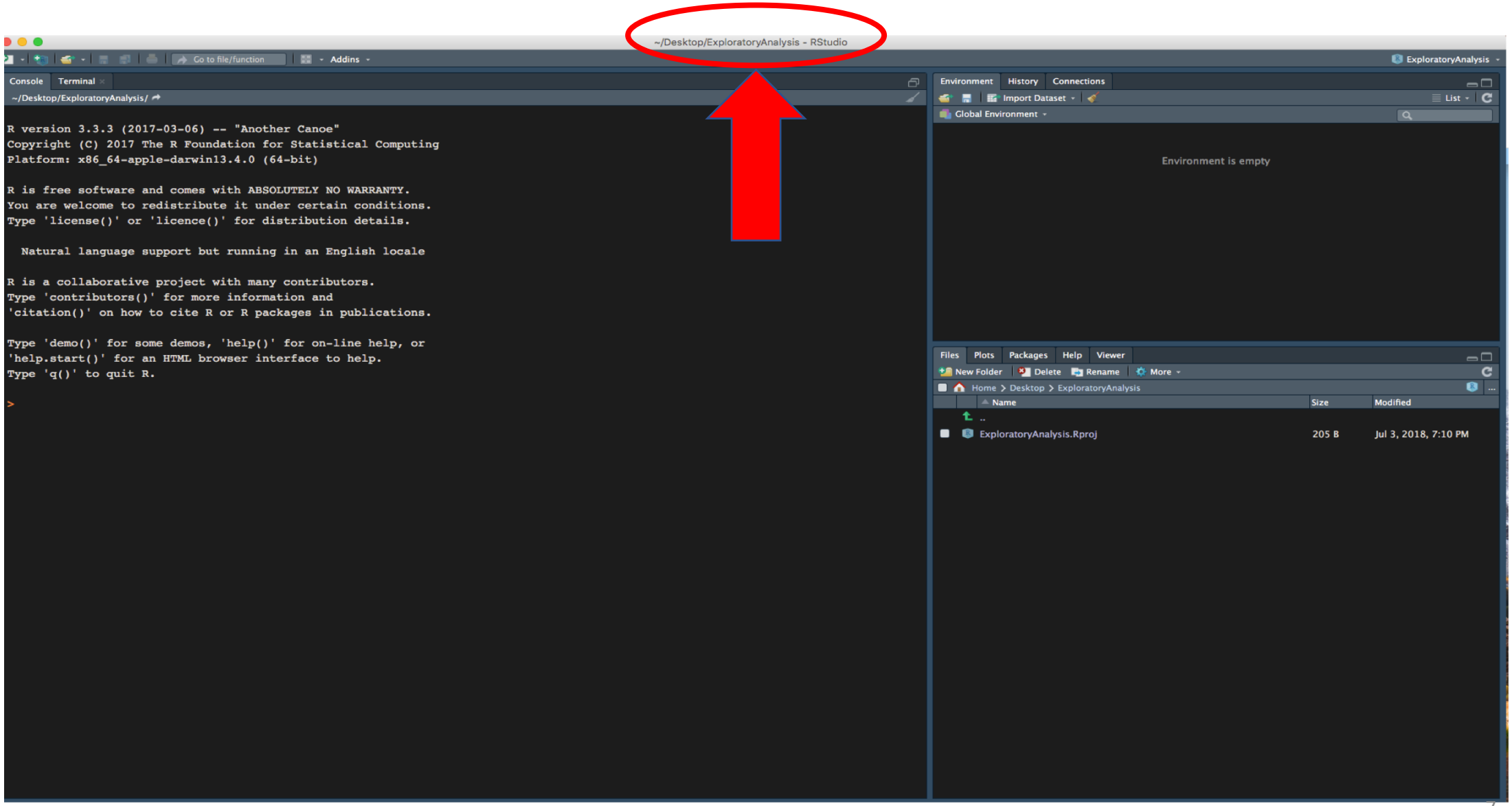
STEP: 2



R project

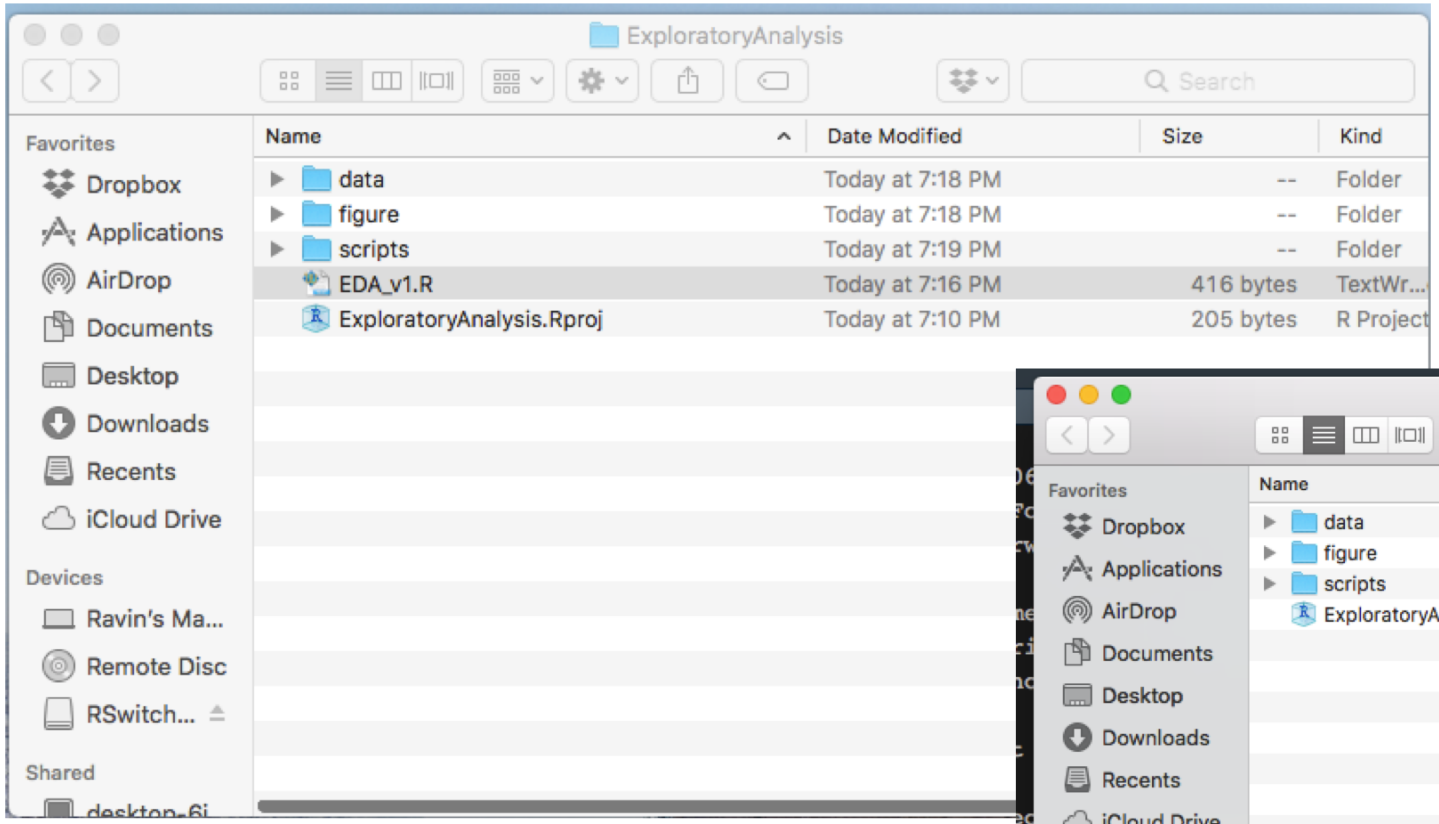

STEP: 3
New Project
Back
Create Project from Existing Directory
Project working directory:
~/Desktop
Browse...
Open in new session
Desktop
Favorites
Recents
Dropbox
Applications
Documents
Desktop
Downloads
iCloud Drive
Devices
Ravin's MacBo...
Remote Disc
RSwitch-1.2
Shared
fiona-inspiron-
hp9709a1
macbookpro-4...
macbookpro-8...
Tags
Name
Date Modified
Size
Kind
ExploratoryAnalysis
Desktop.Rproj
EDA_v1.R
Rplots.pdf
Exploratory Data Analysis.pptx
ARP_Fig3a_last_part
ARP_Fig3a_last_part.pptx
Presentation1.pptx
ARP_Figure3Axx
Epi summer 2018 v2.pptx
VISA_mom
Rplot03
PY56CH25-Garrett[001-022] copy
ARP_Figure3A.png
2012 CERES proposal v30a.docx
Garrett_et_al_SARE_RE_12 11 06_final.docx
New Folder
Cancel
Open
STEP: 4
New Project
Back
Create Project from Existing Directory
Project working directory:
~/Desktop/ExploratoryAnalysis
Browse...
STEP: 5
Open in new session
Create Project
Cancel

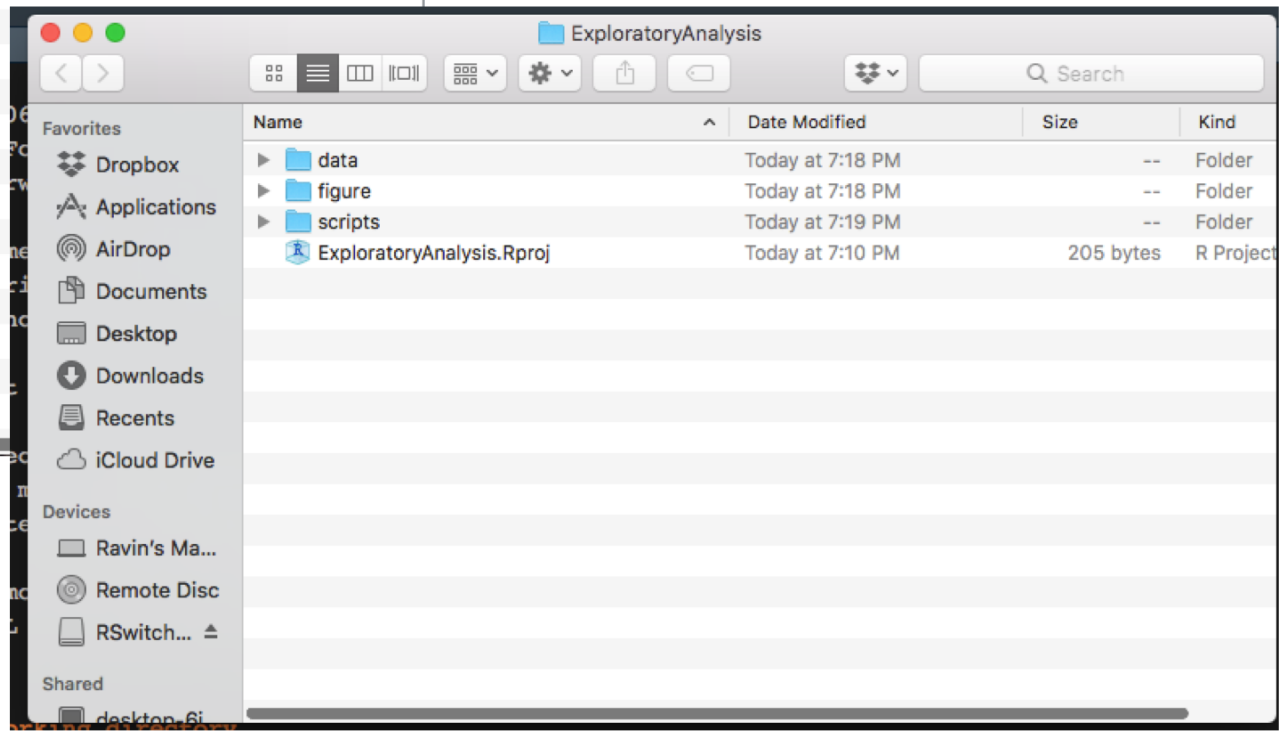
R project


~/Desktop/ExploratoryAnalysis - RStudio
Go to file/function
Addins
ExploratoryAnalysis
Console
Terminal
~/Desktop/ExploratoryAnalysis/
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.
Natural language support but running in an English locale
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.
>
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home > Desktop > ExploratoryAnalysis
Name
Size
Modified
..
ExploratoryAnalysis.Rproj
205 B
Jul 3, 2018, 7:10 PM

R project

Everything you need is in one place, and cleanly separated into sub-folders





Advantages

- Allow to save all materials related to a single analysis in one working environment and sub-folders
- No need to worry about the file paths - less error
- Easy sharing and reproducible
- Saving working environment and output objects save time, especially if your input file is too large

Part Two: Exploratory Data Analysis

Exploratory Data Analysis: How we dissect a data set; what we look for; how we look; and how we interpret

- Generate questions about your data.
- Search for answers by visualizing, transforming, and modelling your data.
- Use what you learn to refine your questions and/or generate new questions.

Advantages:

- maximize insight into a data set
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- develop models

Exploratory Data Analysis

Diamonds: Built in dataset in R

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

Exploratory Data Analysis