



Exploratory Data Analysis

By
Ravin Poudel
Garrett Lab

Summer 2018
Epidemiology and Data Science



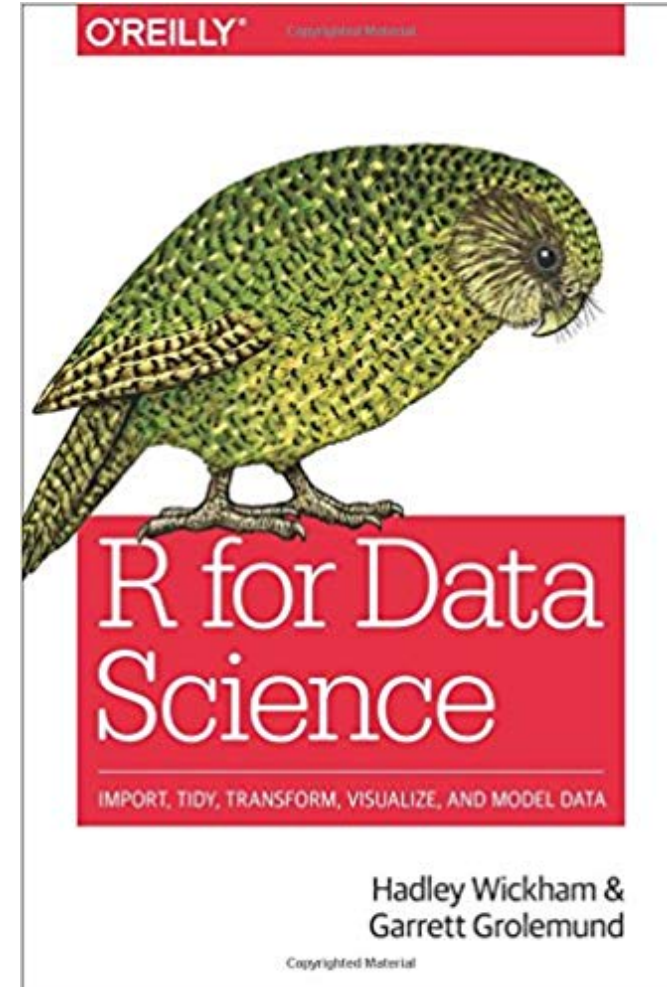
Learning objectives

- 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

The screenshot displays the RStudio environment with the following components:

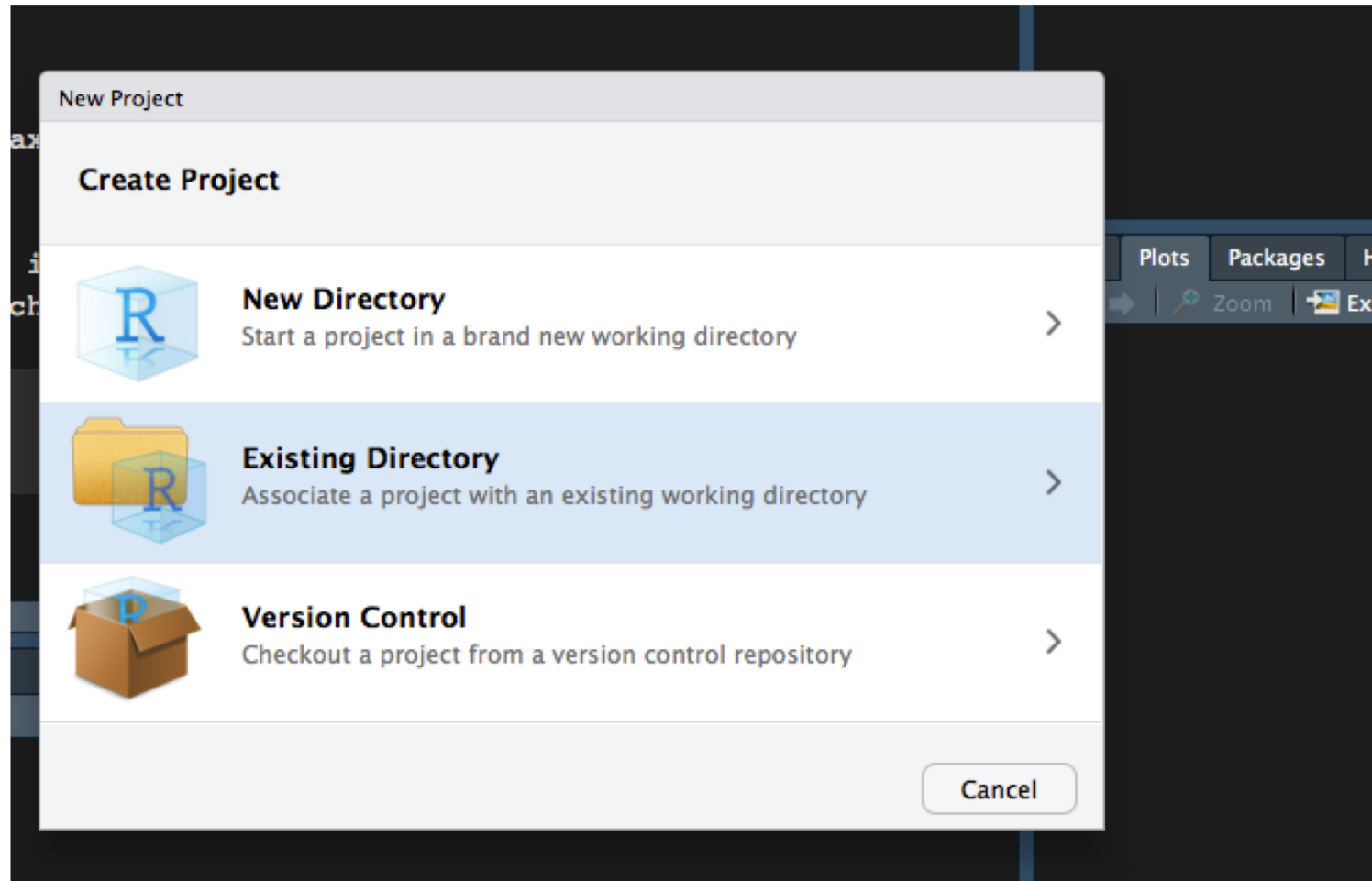
- Menu:** A dropdown menu is open, showing options like "New File", "New Project...", "Open File...", "Save", "Knit Document", and "Quit Session...".
- Environment:** The "Global Environment" is empty, displaying "Environment is empty".
- Console:** The terminal shows the following R code and output:

```
1 1 (Top Level) R Script
~/Desktop/ ↗
9      (3.75,4.25]    4
10     (4.25,4.75]    1
11     (4.75,5.25]    1
> diamonds %>%
+   count(cut_width(carat, 4))
# A tibble: 2 x 2
  `cut_width(carat, 4)`    n
      <fctr> <int>
1      [-2,2] 52051
2      (2,6] 1889
> ggplot(diamonds) +
+   geom_histogram(mapping = aes(x = y), binwidth = 0.5) +
+   coord_cartesian(ylim = c(0, 50))
# 7.5.1 A categorical and continuous variable
> ggplot(data = diamonds, mapping = aes(x = price)) +
+   geom_freqpoly(mapping = aes(colour = cut), binwidth = 500)
>
```
- Plots:** A frequency polygon plot is shown in the "Plots" pane. The x-axis is labeled "price" (ranging from 0 to 20,000) and the y-axis is labeled "count" (ranging from 0 to 5,000). The plot displays five lines representing different diamond cuts: Fair (red), Good (green), Very Good (blue), Premium (cyan), and Ideal (magenta). The "Ideal" cut shows the highest frequency, peaking at approximately 5,000 counts for prices around 1,000.



Part One: R project

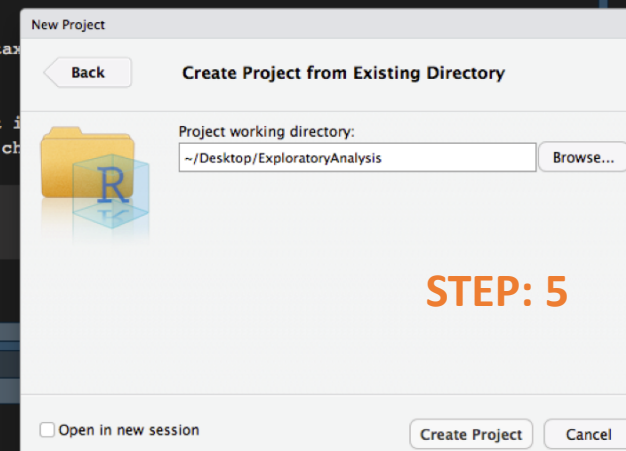
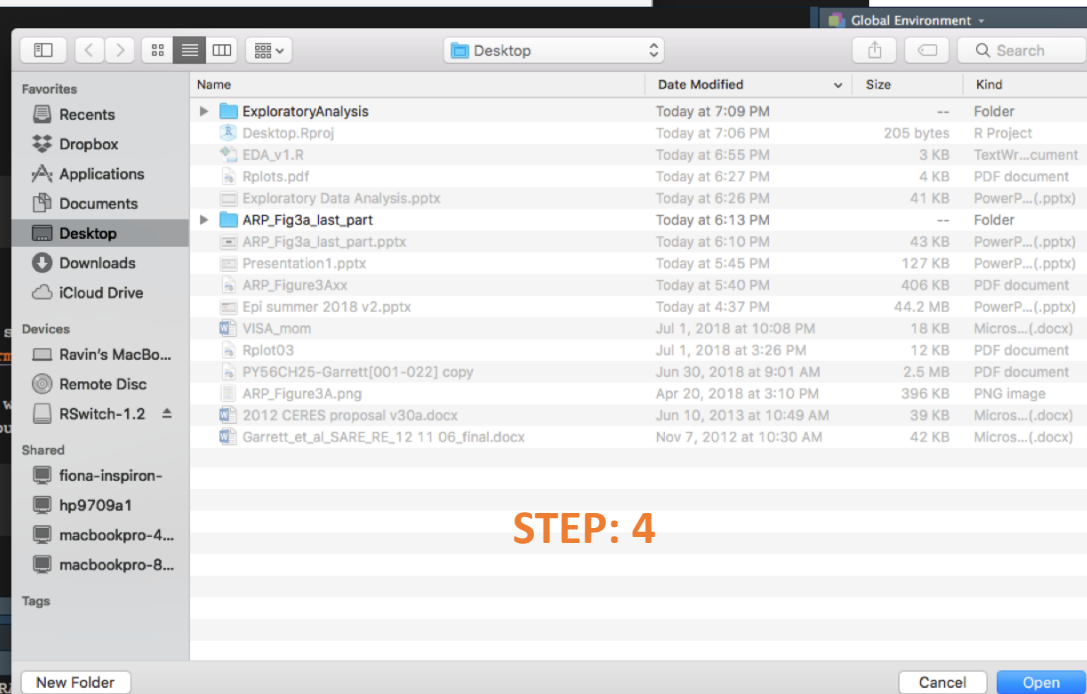
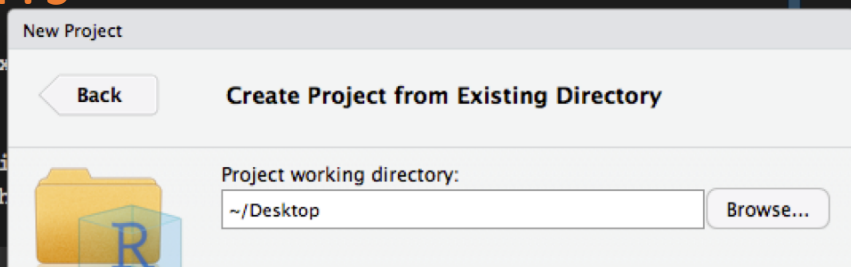
STEP: 2





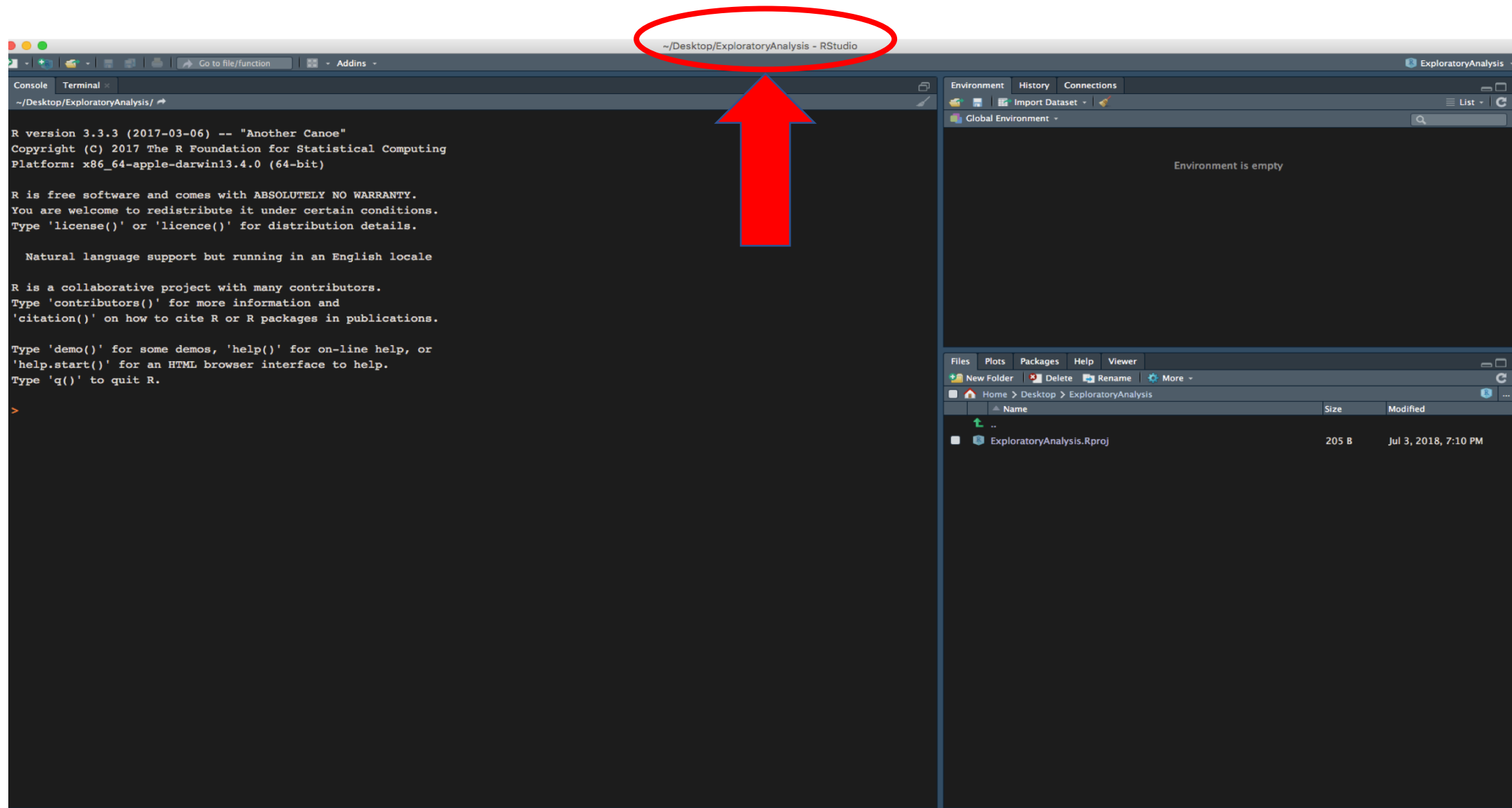
R project

STEP: 3





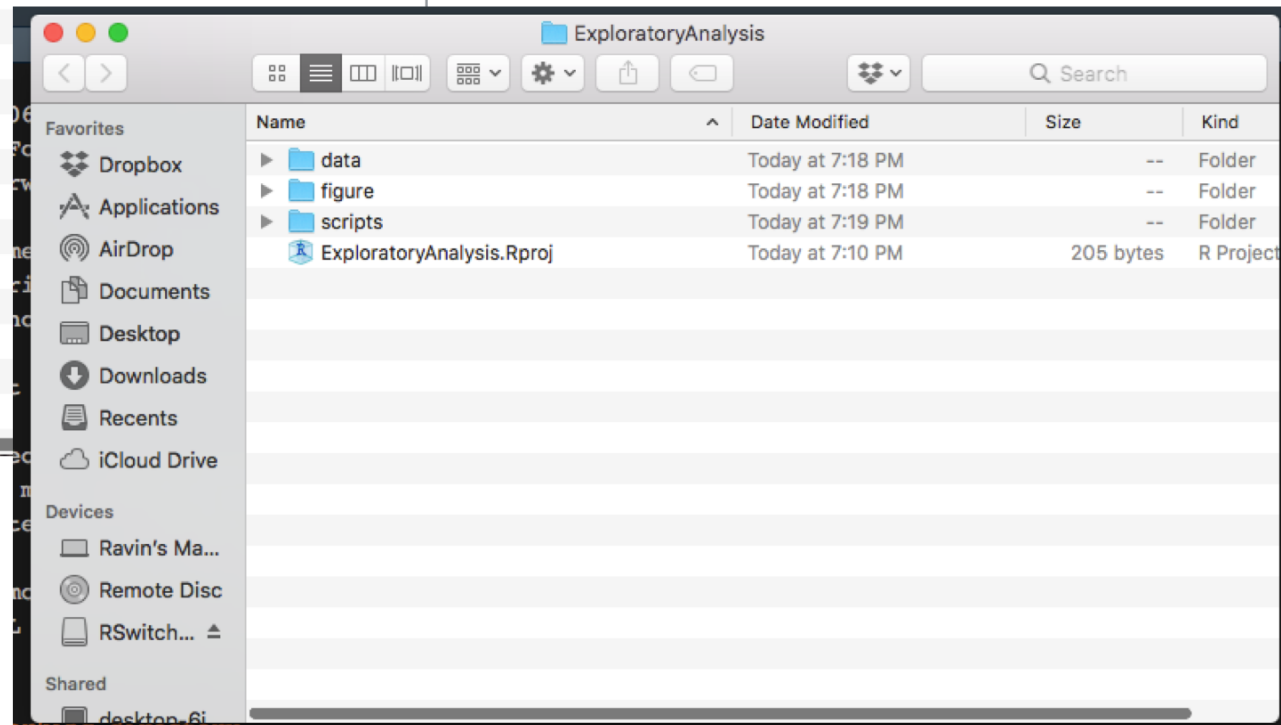
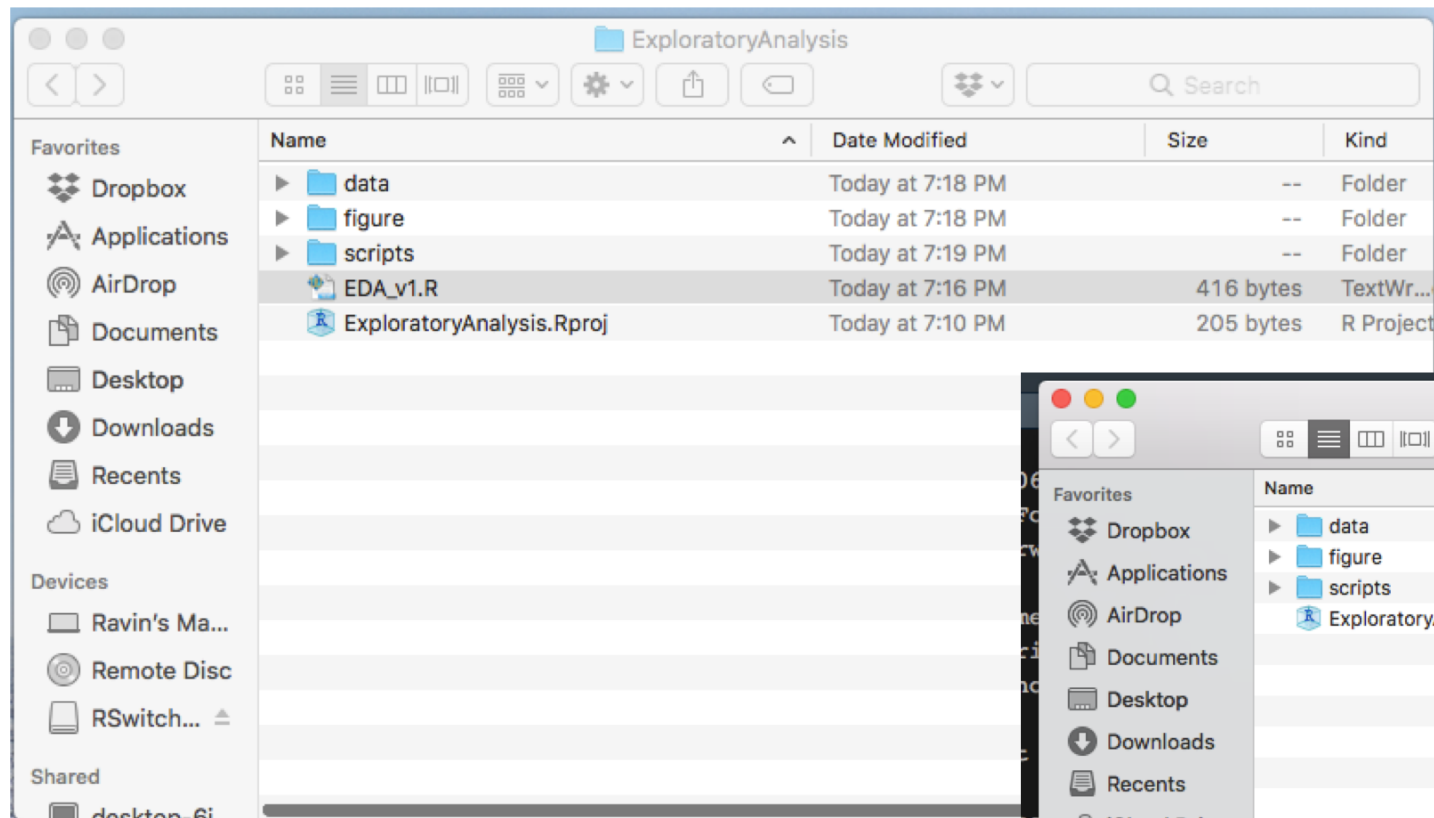
R project





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 = $z / \text{mean}(x, y) = 2 * z / (x + y)$ (43--79)
table	width of top of diamond relative to widest point (43--95)



Exploratory Data Analysis

