Introduction to Reproducible Research in R and R Studio.

Susan Johnston

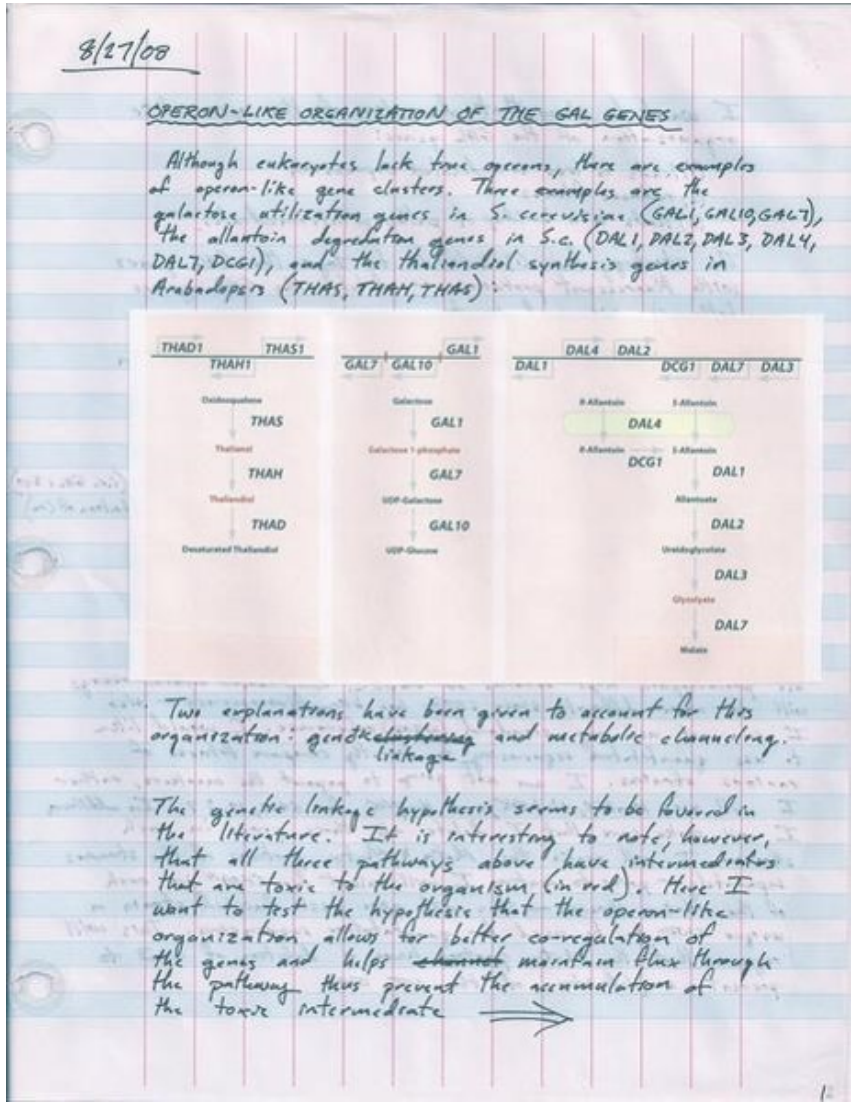
April 1, 2016

What is Reproducible Research?

Reproducibility *is the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else working independently, [and] is one of the main principles of the scientific method.*

Wikipedia

In the lab:



8/27/08

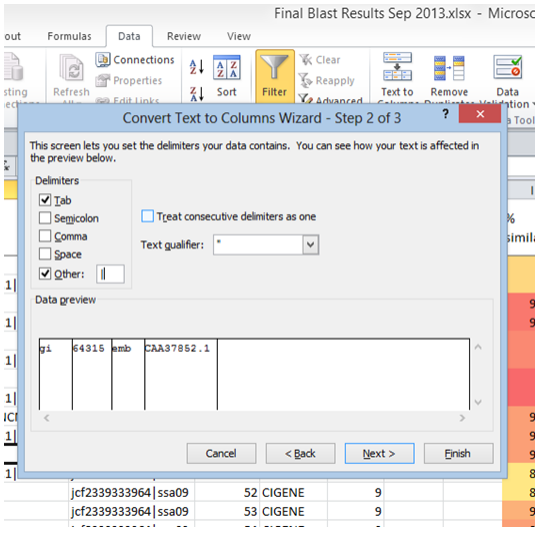
OPERON-LIKE ORGANIZATION OF THE GAL GENES

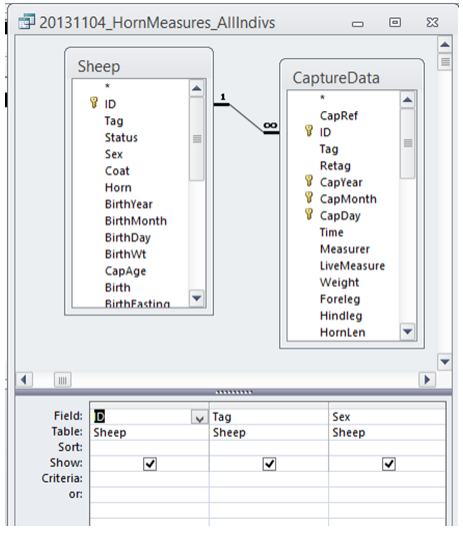
Although eukaryotes lack true operons, there are examples of operon-like gene clusters. Three examples are the galactose utilization genes in S. cerevisiae (GAL1, GAL10, GAL7), the allantoin degradation genes in S.c. (DAL1, DAL2, DAL3, DAL4, DAL7, DCG1), and the thalianadiol synthesis genes in Arabadopsis (THAS, THAH, THAD):

Two explanations have been given to account for this organization: genetic ~~clustering~~ linkage and metabolic channeling.

The genetic linkage hypothesis seems to be favored in the literature. It is interesting to note, however, that all three pathways above have intermediates that are toxic to the organism (in red). Here I want to test the hypothesis that the operon-like organization allows for better co-regulation of the genes and helps ~~channel~~ maintain flux through the pathway thus prevent the accumulation of the toxic intermediate →

Many of us are clicking, copying and pasting...





Haggis population density in the Scottish Highlands

S Johnston, University of Edinburgh.

Introduction.

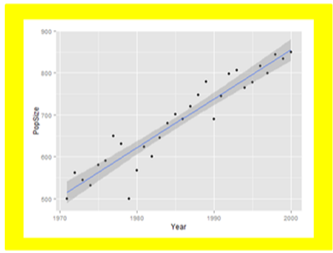


Figure 1: Linear regression of haggis population size and year.

The haggis is a common pest species in the Scottish Highlands. Haggis population densities were recorded annually from 1970 to 2000. We found that the haggis population size increased over this period by 11.67 individuals year^{-1} ($P < 0.001$, Figure 1).

- Can you repeat all of this again...
- ...and would you get the same results every time?

Worst Case Scenario

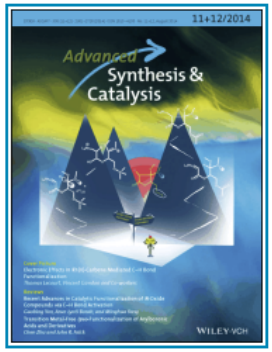
Retraction Watch

Tracking retractions as a window into the scientific process

Archive for the 'not reproducible' Category

Molecular mixup burns chemistry paper

without comments



Chemists at Lanzhou University in China did the right thing last month, retracting a paper in *Advanced Synthesis & Catalysis* because of issues with a reactant that could only be corrected by changing "all the text and quantities."

When the scientists were adding what was labeled Reactant 1 to the mix, they believed it was α-ethoxycarbonyl-α-azido-*N*-phenylacetamides. Unfortunately, what they were actually using was a decomposed version of the molecule, which threw everything off.

Here's the notice for "*tert*-Butyl Hydroperoxide and Tetrabutylammonium Iodide- Promoted Free Radical Cyclization of α-Azido-*N*-arylamides": Read the rest of this entry »

Share this:

Email | Facebook 49 | Twitter 12

Written by Cat Ferguson
April 14th, 2015 at 11:30 am

Posted in Advanced Synthesis and Catalysis,chemistry retractions,china retractions,doing the right thing,freely available,not reproducible,wiley

Two more retractions bring lab break-in biochemist up to eleven

without comments

Karel Bezouška, the Czech biochemist who was caught on hidden camera breaking into a lab fridge to fake results, has turned it up to eleven with two new retractions.

Both retractions appeared in *Biochemical and Biophysical Research Communications,* one in October 2014 and one in January 2015. His story began two decades ago in 1994, when he published a paper in *Nature* that couldn't be reproduced, and was eventually retracted in 2013.

The best part of the story, of course, is that when his university was attempting to recreate his experiments, Bezouška broke into a lab fridge to tamper with the experiments. Unbeknownst to him, he was caught on hidden camera. Read the rest of this entry »

Share this:

Subscribe to Blog

Email Address
Subscribe

Pages

How you can support Retraction Watch
Meet the Retraction Watch staff
About Adam Marcus
About Ivan Oransky
The Center For Scientific Integrity
Board of Directors
The Retraction Watch FAQ, including comments policy
The Retraction Watch Transparency Index
The Retraction Watch Store
Upcoming Retraction Watch appearances
What people are saying about Retraction Watch

Search for:
Search

Recent Posts

Beleaguered Förster turns down prestigious professorship, citing personal toll

We're on Facebook

Retraction Watch
Like 3,135

Scenarios that benefit from reproducibility

- New raw data becomes available.
- You return to the project after a period of time.
- Project gets handed to new PhD student/postdoc.
- Working collaboratively.
- A reviewer wants you to change a model parameter.
- When you find an error, but not sure where you went wrong.

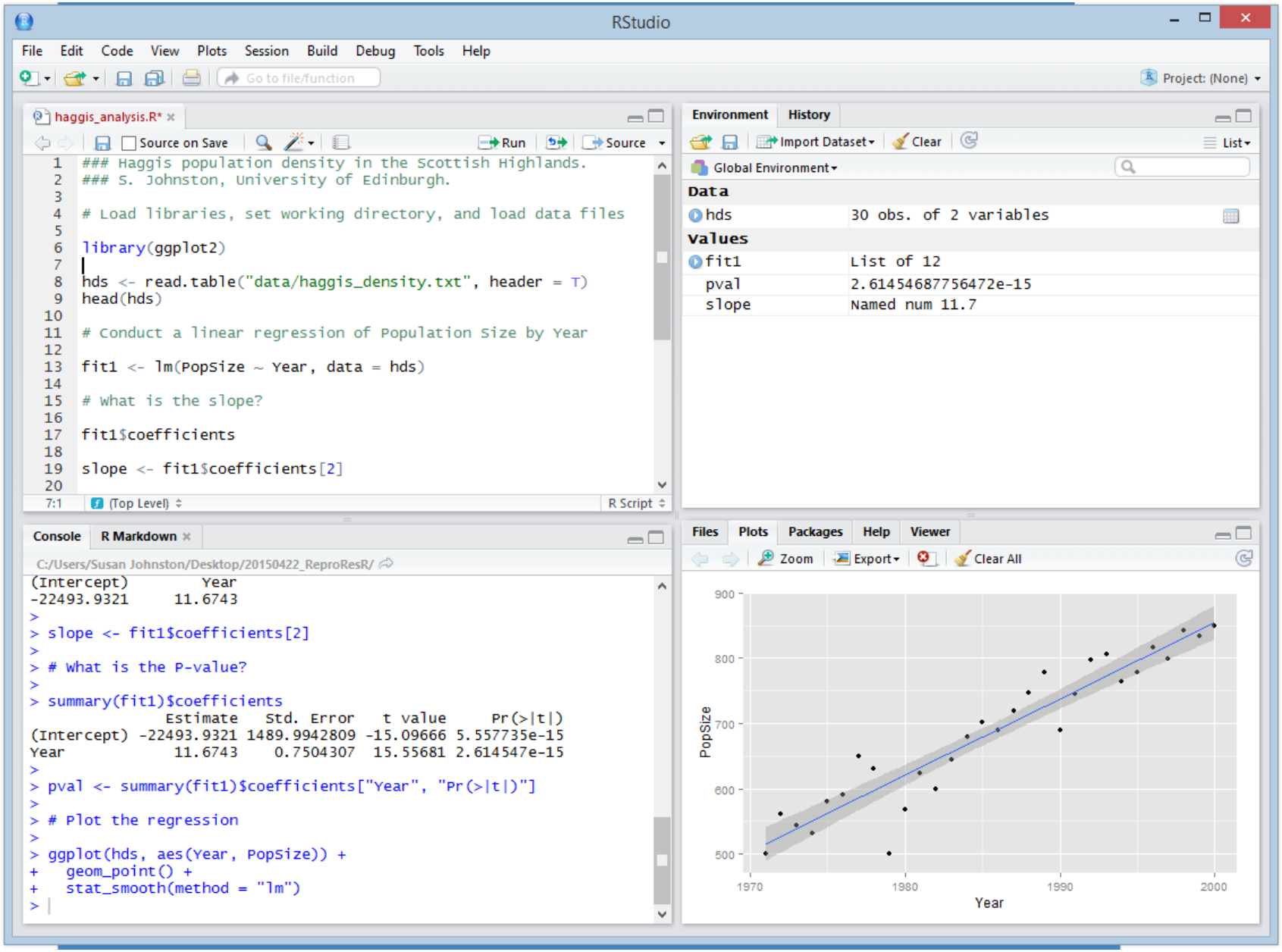
Four rules for reproducibility.

1. Create a portable project.
2. Avoid manual data manipulation steps - **use code!**
3. Connect results to text.
4. Version control all custom scripts and documents.

Disclaimer

Many solutions to the same problem!

The RStudio Environment: http://www.rstudio.com



Reproducible Research in RStudio

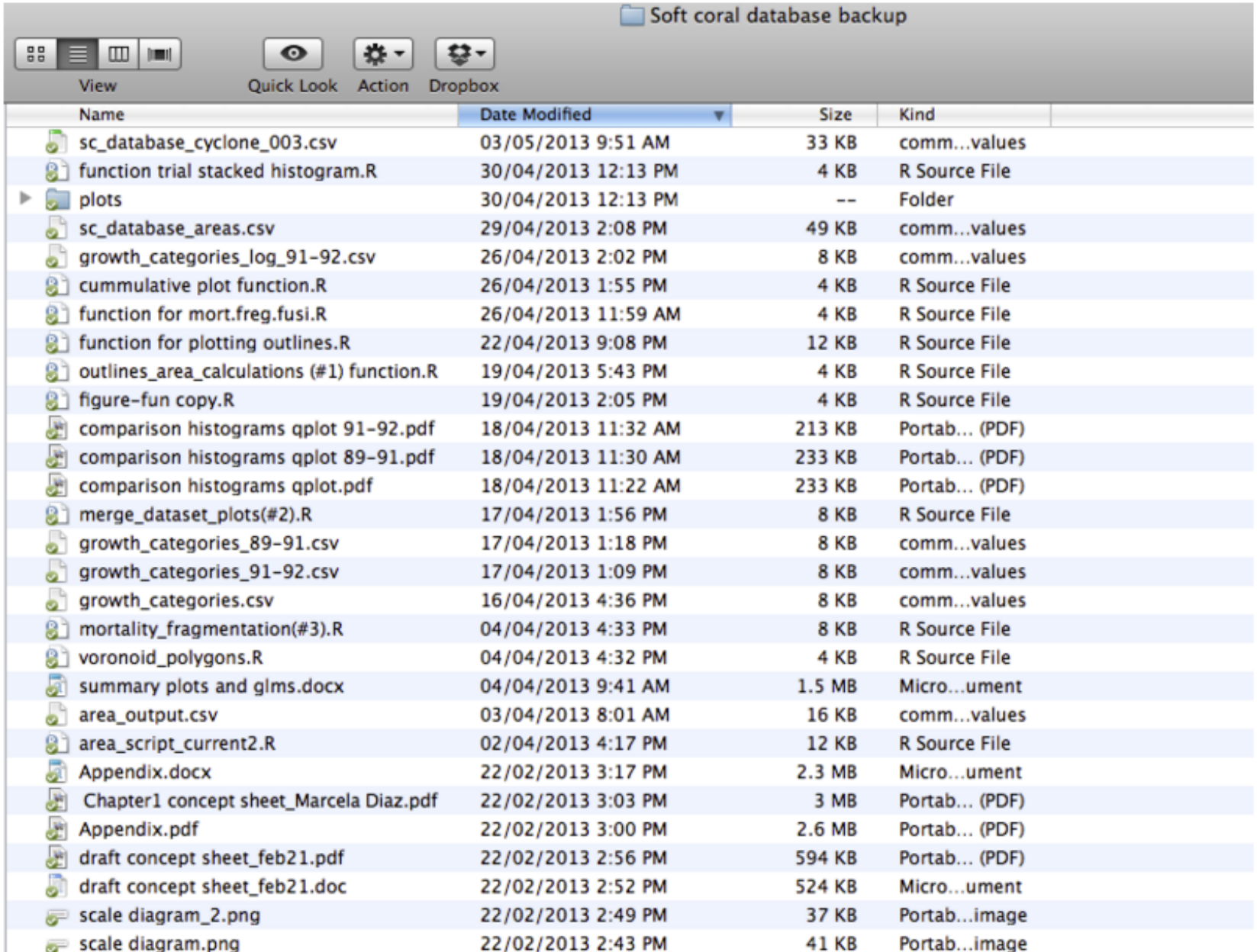
1. Creating a Portable Project (.Rproj)
2. Automate analyses - stop clicking and start typing.
3. Dynamic report writing with R Markdown and `knitr`
4. Version control using `git`

Reproducible Research in RStudio.

1. Creating a Portable Project (.Rproj)
2. Automate analyses - stop clicking and start typing.
3. Dynamic report writing with R Markdown and `knitr`
4. Version control using `git`

Structuring an R Project.

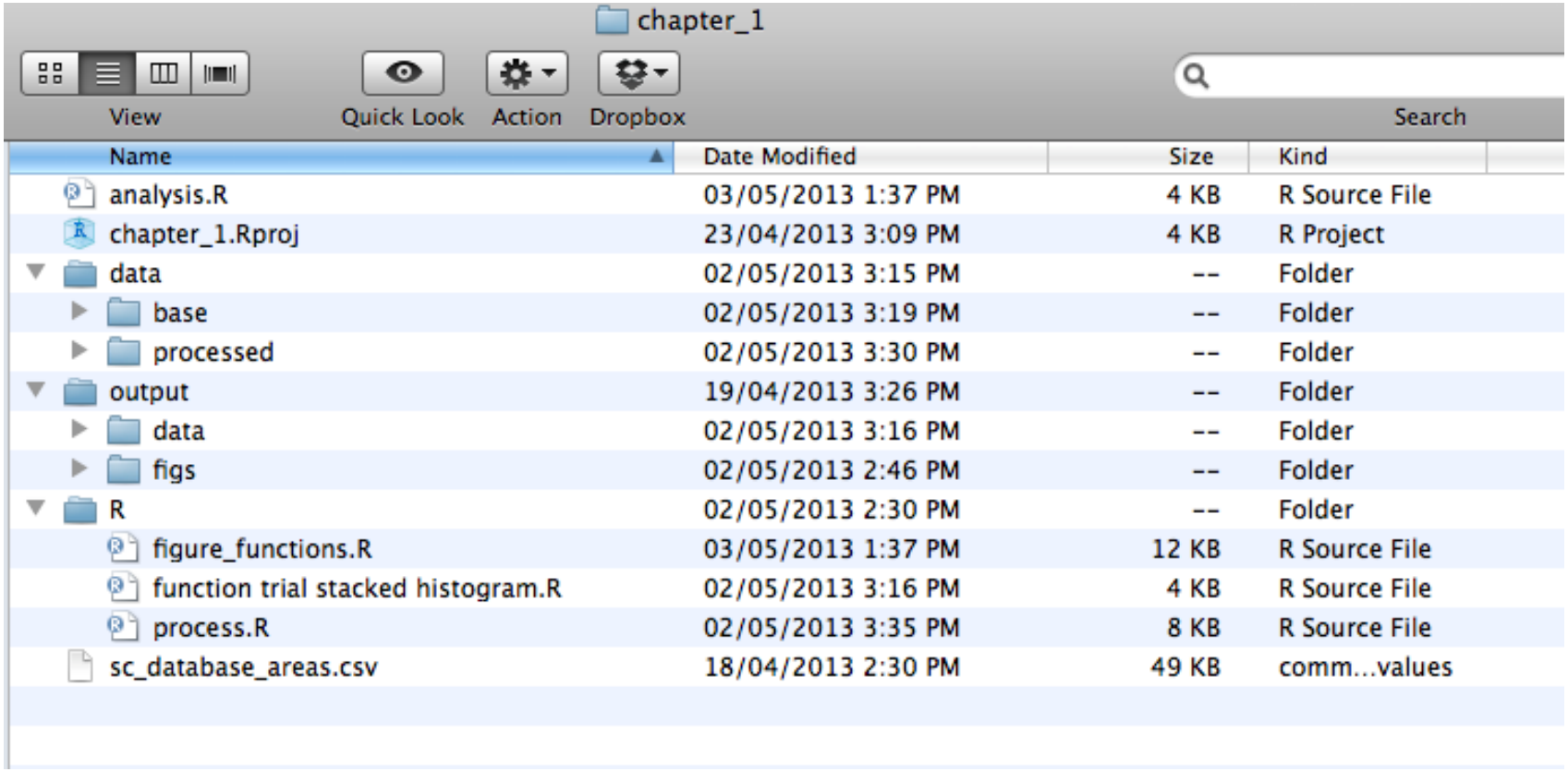
http://nicercode.github.io/blog/2013-05-17-organising-my-project/



Name	Date Modified	Size	Kind
sc_database_cyclone_003.csv	03/05/2013 9:51 AM	33 KB	comm...values
function trial stacked histogram.R	30/04/2013 12:13 PM	4 KB	R Source File
plots	30/04/2013 12:13 PM	--	Folder
sc_database_areas.csv	29/04/2013 2:08 PM	49 KB	comm...values
growth_categories_log_91-92.csv	26/04/2013 2:02 PM	8 KB	comm...values
cummulative plot function.R	26/04/2013 1:55 PM	4 KB	R Source File
function for mort.freg.fusi.R	26/04/2013 11:59 AM	4 KB	R Source File
function for plotting outlines.R	22/04/2013 9:08 PM	12 KB	R Source File
outlines_area_calculations (#1) function.R	19/04/2013 5:43 PM	4 KB	R Source File
figure-fun copy.R	19/04/2013 2:05 PM	4 KB	R Source File
comparison histograms qplot 91-92.pdf	18/04/2013 11:32 AM	213 KB	Portab... (PDF)
comparison histograms qplot 89-91.pdf	18/04/2013 11:30 AM	233 KB	Portab... (PDF)
comparison histograms qplot.pdf	18/04/2013 11:22 AM	233 KB	Portab... (PDF)
merge_dataset_plots(#2).R	17/04/2013 1:56 PM	8 KB	R Source File
growth_categories_89-91.csv	17/04/2013 1:18 PM	8 KB	comm...values
growth_categories_91-92.csv	17/04/2013 1:09 PM	8 KB	comm...values
growth_categories.csv	16/04/2013 4:36 PM	8 KB	comm...values
mortality_fragmentation(#3).R	04/04/2013 4:33 PM	8 KB	R Source File
voronoid_polygons.R	04/04/2013 4:32 PM	4 KB	R Source File
summary plots and glms.docx	04/04/2013 9:41 AM	1.5 MB	Micro...ument
area_output.csv	03/04/2013 8:01 AM	16 KB	comm...values
area_script_current2.R	02/04/2013 4:17 PM	12 KB	R Source File
Appendix.docx	22/02/2013 3:17 PM	2.3 MB	Micro...ument
Chapter1 concept sheet_Marcela Diaz.pdf	22/02/2013 3:03 PM	3 MB	Portab... (PDF)
Appendix.pdf	22/02/2013 3:00 PM	2.6 MB	Portab... (PDF)
draft concept sheet_feb21.pdf	22/02/2013 2:56 PM	594 KB	Portab... (PDF)
draft concept sheet_feb21.doc	22/02/2013 2:52 PM	524 KB	Micro...ument
scale diagram_2.png	22/02/2013 2:49 PM	37 KB	Portab...image
scale diagram.png	22/02/2013 2:43 PM	41 KB	Portab...image

Structuring an R Project.

http://nicercode.github.io/blog/2013-05-17-organising-my-project/



All data, scripts and output should be kept within the same project directory.

Structuring an R Project.

http://nicercode.github.io/blog/2013-04-05-projects/

R/ Contains functions relevant to analysis.

data/ Contains raw data as **read only**.

doc/ Contains the paper.

figs/ Contains the figures.

output/ Contains analysis output
(processed data, logs, etc. Treat as **disposable**).

.R Code for the analysis.

Structuring an R Project.

http://robjhyndman.com/hyndsight/workflow-in-r/

- **load.R** - read in data from files
- **clean.R** - pre-processing and cleaning
- **functions.R** - define what you need for anlaysis
- **do.R** - do the analysis!

Bad habits can hinder portability.

https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects

Hadley Wickham @hadleywickham · Jan 27

Never plan on sharing your code? Using **setwd**() means that old code will break if you ever reorganise your directories

7 6

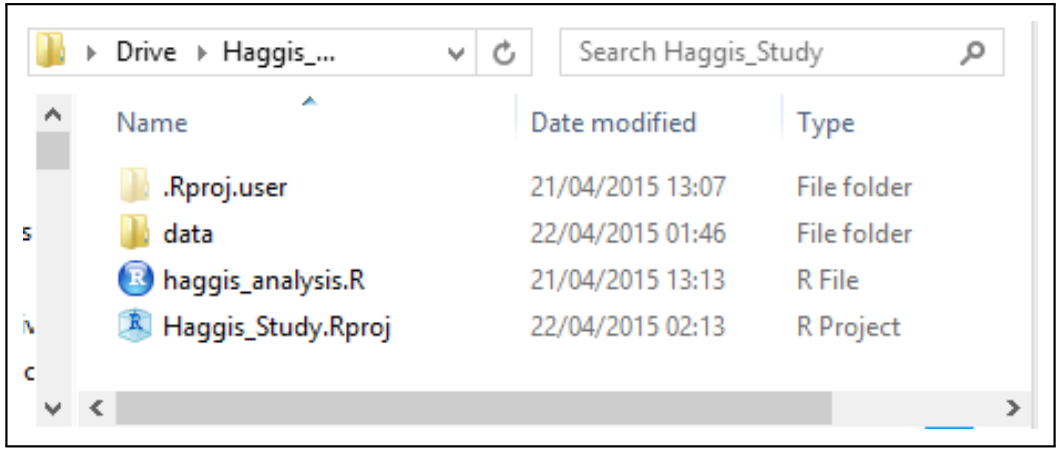
```
setwd("C:/Users/susjoh/Desktop/SalmoAnalysis")
setwd("C:/Users/Susan Johnston/Desktop/SalmoAnalysis")
setwd("C:/Users/sjohns10/Drive/SalmoAnalysis")

source("../../OvisAnalysis/GWASplotfunctions.R")
```

An analysis should be contained within a directory, and it should be easy to move it or pass on to someone new.

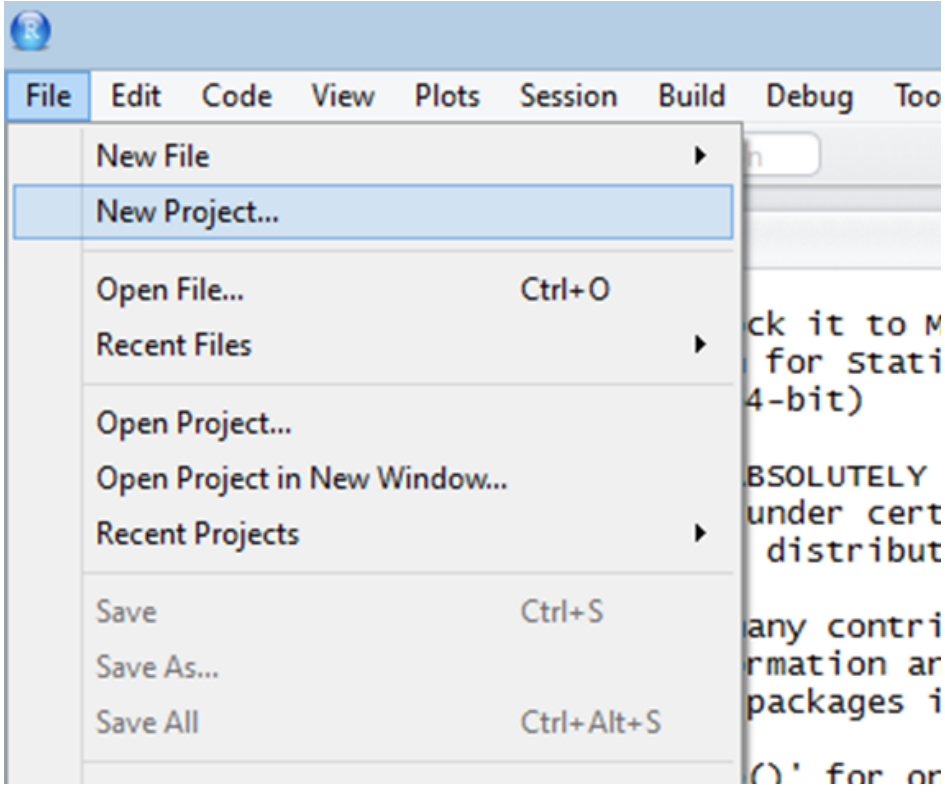
Solution: using RStudio Projects.

https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects

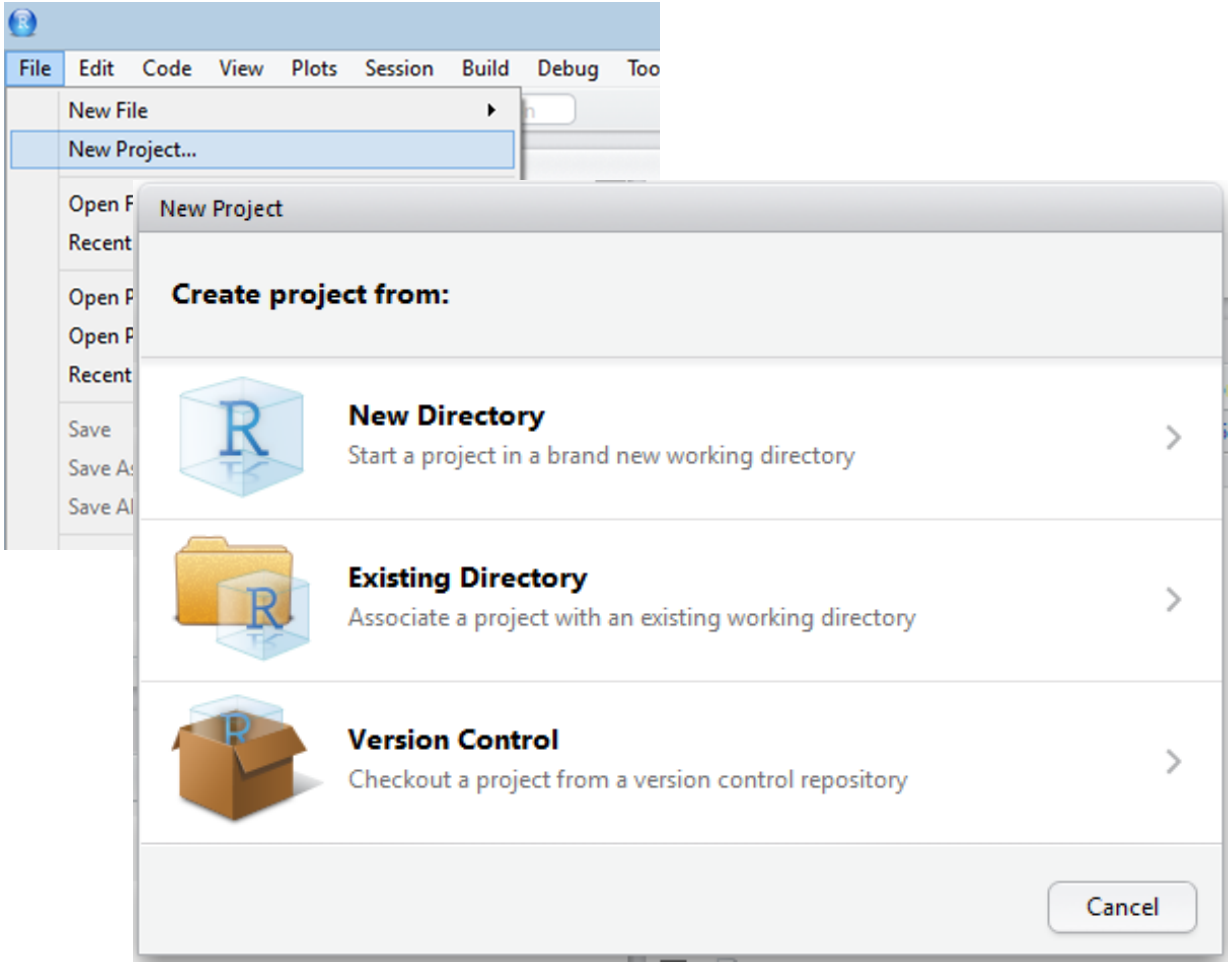


- Establishes a directory with associated `.Rproj` file.
- Automatically sets the working directory.
- Can save and source `.Rprofile`, `.Rhistory`, `.Rdata` files.
- Allows version control within R Studio.

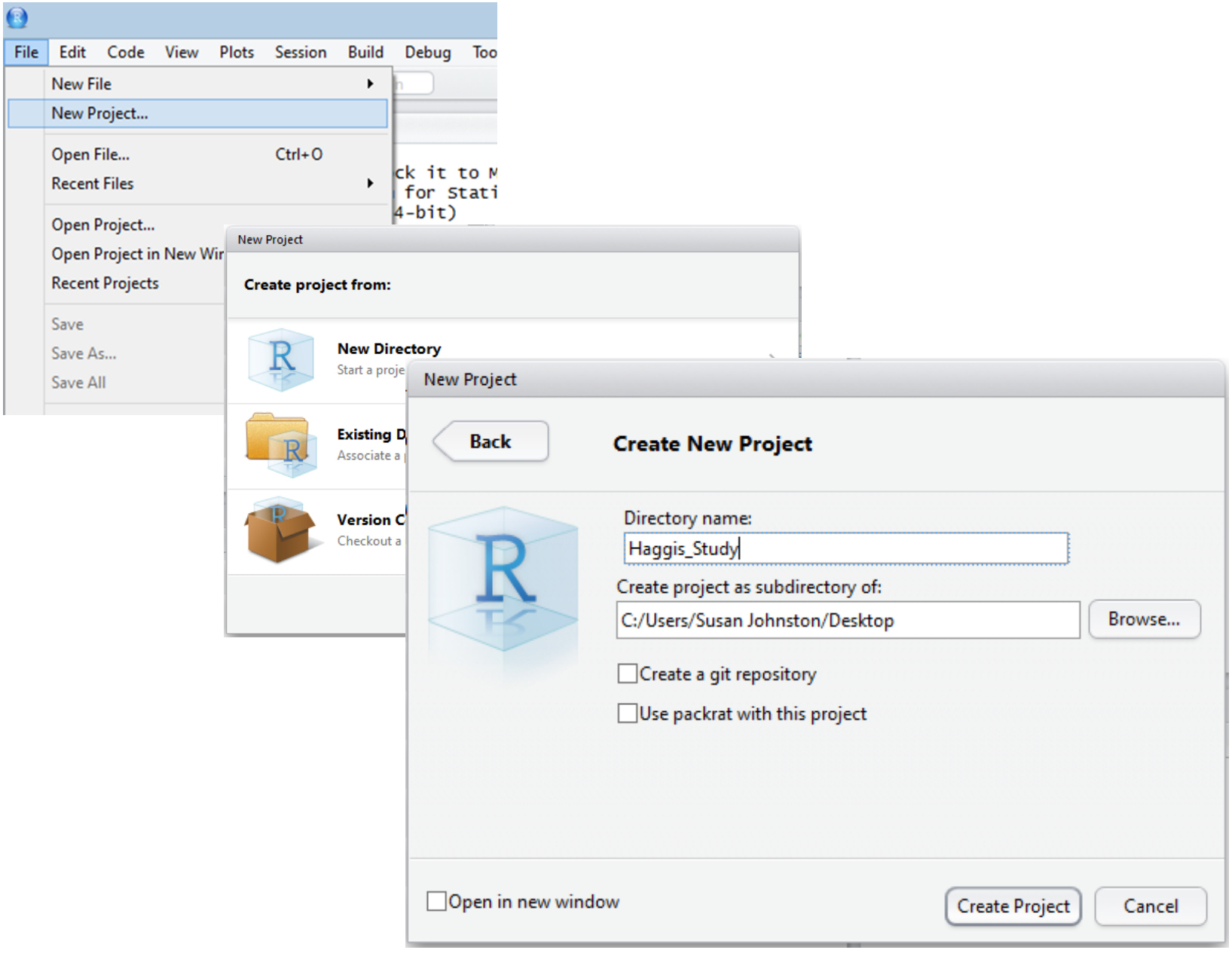
Creating a Portable Project (.Rproj)



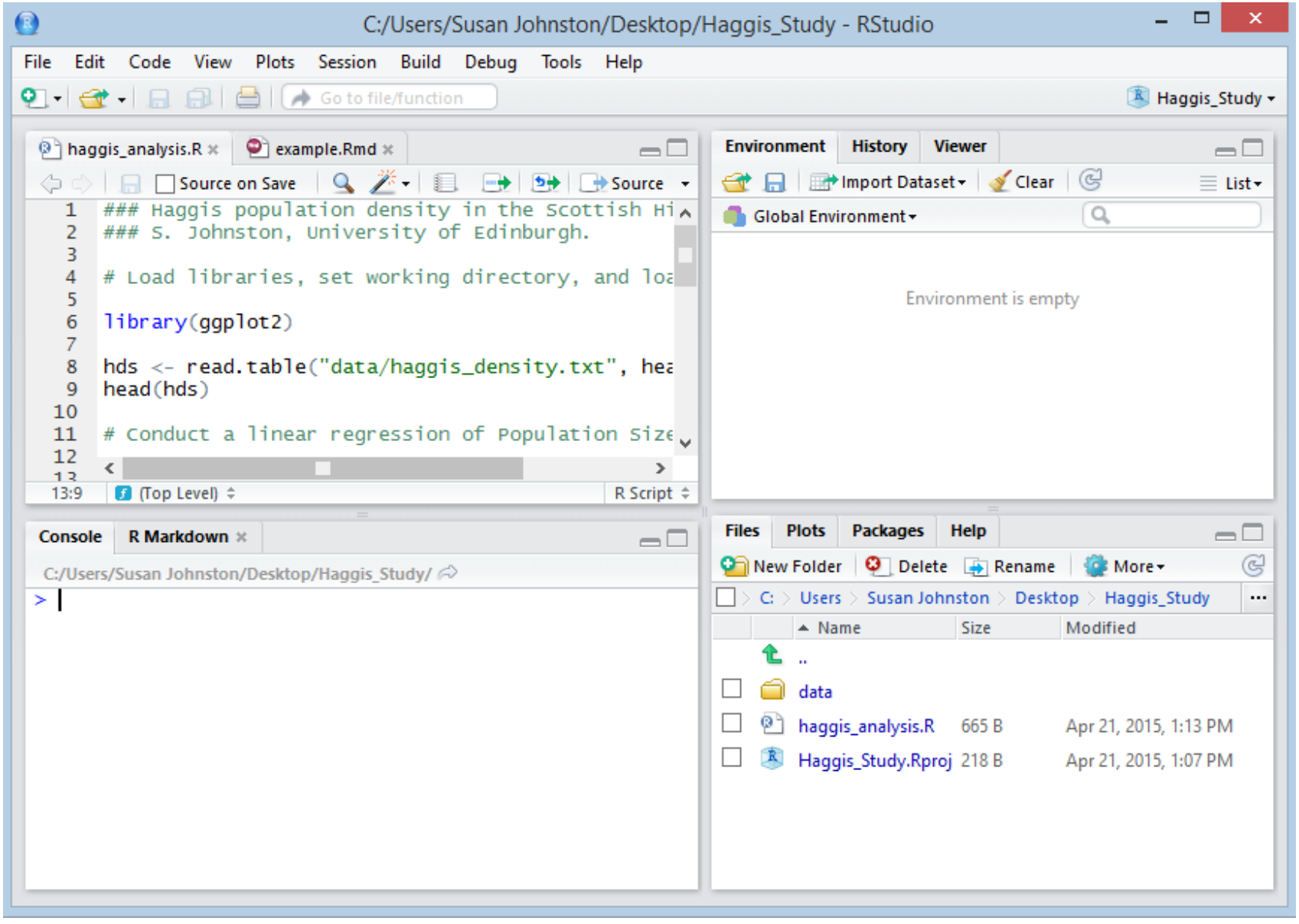
Creating a Portable Project (.Rproj)



Creating a Portable Project (.Rproj)



Creating a Portable Project (.Rproj)



Reproducible Research in RStudio.

1. Creating a Portable Project (.Rproj)
2. Automate analyses - stop clicking and start typing.
3. Dynamic report writing with R Markdown and `knitr`
4. Version control using `git`

This is R. There is no "if". Only "how".

- CRAN, Bioconductor, github
- Reading in data and functions

```
read.table(), read.csv(), read.xlsx(), source()
```

- Reorganising data

```
reshape, plyr, dplyr
```

- Generate figures

```
plot(), library(ggplot2)
```

- Running external programmes with system()

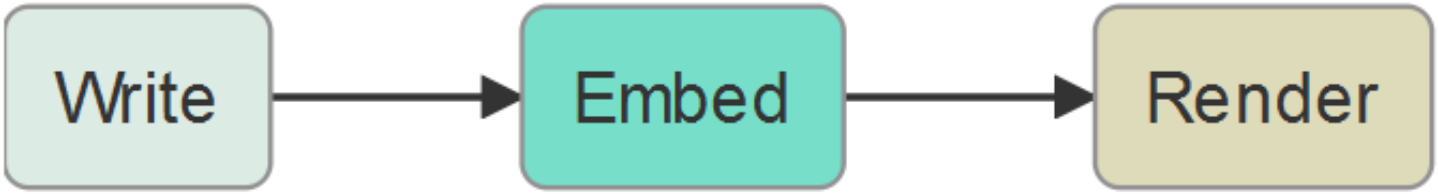
```
Unix/Mac: system("plink -file OvGen --freq")
Windows: system("cmd", input = "plink -file OvGen --freq")
```

Reproducible Research in RStudio.

1. Creating a Portable Project (.Rproj)
2. Automate analyses - stop clicking and start typing.
3. Dynamic report writing with R Markdown and `knitr`
4. Version control using `git`

knitr

Elegant, flexible and fast dynamic report generation with R



The knitr package allows R code and document templates to be compiled into a single report containing text, results and figures.

knitr

Elegant, flexible and fast dynamic report generation with R

Output script as Notebook

Compile notebook

20150422_ReproResR2.tex | haggis_analysis.R

Source on Save | Run | Source

```
1  ### Haggis population density in the Scottish Highlands.
2  ### S. Johnston, University of Edinburgh.
3
4  # Load libraries, set working directory, and load data files
5
6  library(ggplot2)
7
8  hds <- read.table("data/haggis_density.txt", header = T)
9  head(hds)
10
11 # Conduct a linear regression of Population Size by Year
12
13 fit1 <- lm(PopSize ~ Year, data = hds)
14
15 # What is the slope?
16
17 fit1$coefficients
18
19 slope <- fit1$coefficients[2]
20
21 # What is the P-value?
22
23 summary(fit1)$coefficients
24
25 pval <- summary(fit1)$coefficients["Year", "Pr(>|t|)"]
26
27 # Plot the regression
28
29 ggplot(hds, aes(Year, PopSize)) +
30   geom_point() +
31   stat_smooth(method = "lm")
32
33
```

1:1 (Top Level) R Script

haggis_analysis.pdf - [haggis_analysis.R] - SumatraPDF

File View Go To Zoom Settings Help

Page: 1 / 2 Find:

haggis_analysis.R

sjohns10

Fri Apr 01 14:22:50 2016

```
### Haggis population density in the Scottish Highlands.
### S. Johnston, University of Edinburgh.

# Load libraries, set working directory, and load data files

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
hds <- read.table("data/haggis_density.txt", header = T)
head(hds)
```

```
##   Year PopSize
## 1 1971     500
## 2 1972     562
## 3 1973     544
## 4 1974     532
## 5 1975     580
## 6 1976     590
```

```
# Conduct a linear regression of Population Size by Year

fit1 <- lm(PopSize ~ Year, data = hds)

# What is the slope?

fit1$coefficients
```

```
## (Intercept)        Year
## -22493.9321     11.6743
```

knitr

Elegant, flexible and fast
dynamic report generation with R

Write reports directly in R

haggis_analysis.R | example.Rmd

Knit HTML | Chunks

```
1  ### Example of using "R Markdown" and `knitr`
2
3  This is an R Markdown Document. Markdown is a
   simple formatting syntax for authoring
   documents. For example, text can be emphasise
   d with *italics* and **bold**.
4
5  Wrap code to be evaluated in R in an R
   "chunk"
6
7  ```{r}
8  nrow(cars)
9  ```
10
11 Code can be evaulated in-line using backticks
   . For example, we can write 1 + 1 = `r 1+1`.
12
13 Plots will also be printed.
14
15 ```{r fig.width = 3, fig.height = 3}
16 library(ggplot2)
17 ggplot(cars, aes(speed, dist)) +
18   geom_point()
19 ```
20
```

8:11 | Chunk 1 | R Markdown

knitr

Elegant, flexible and fast dynamic report generation with R

Write reports directly in R

C:/Users/Susan Johnston/Desktop/Drive/Haggis_Study/example.html

example.html | Open in Browser | Publish | Find

Example of using "R Markdown" and `knitr`

This is an R Markdown Document. Markdown is a simple formatting syntax for authoring documents. For example, text can be emphasised with *italics* and **bold**.

Wrap code to be evaluated in R in an R "chunk"

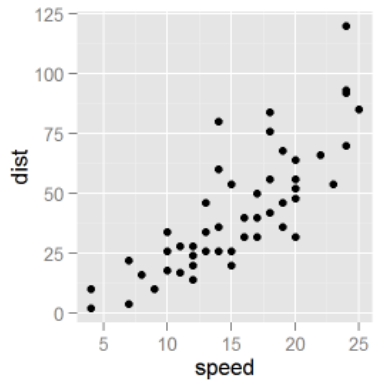
```
nrow(cars)
```

```
## [1] 50
```

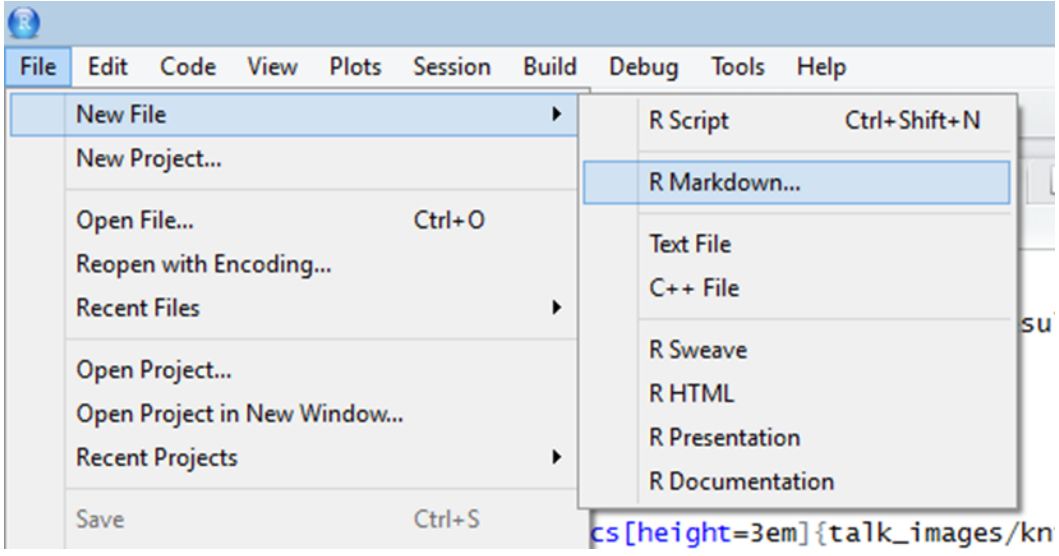
Code can be evaulated in-line using backticks. For example, we can write 1 + 1 = 2.

Plots will also be printed.

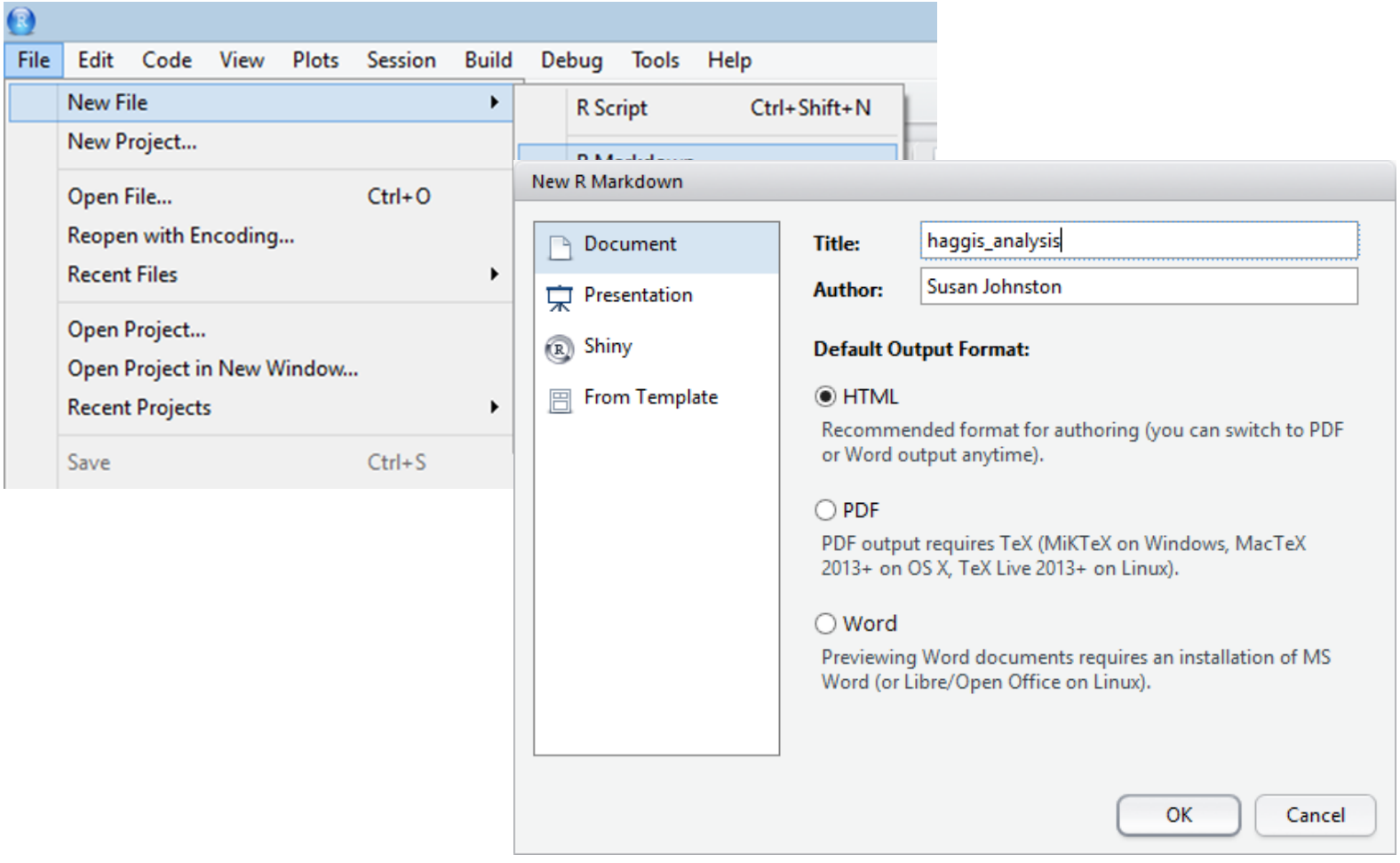
```
library(ggplot2)
ggplot(cars, aes(speed, dist)) +
  geom_point()
```



Creating an R Markdown Script (.Rmd).



Creating an R Markdown Script (.Rmd).



A Quick Start Guide

http://nicercode.github.io/guides/reports/

1. Type report text into .Rmd file

```
Lorem ipsum dolor sit amet, consectetuer adipiscing elit.
```

2. Enclose code to be evaluated in chunks

````
```{r}
model1 <- lm(speed ~ dist, data = cars)
```
````

3. Evaluate code inline

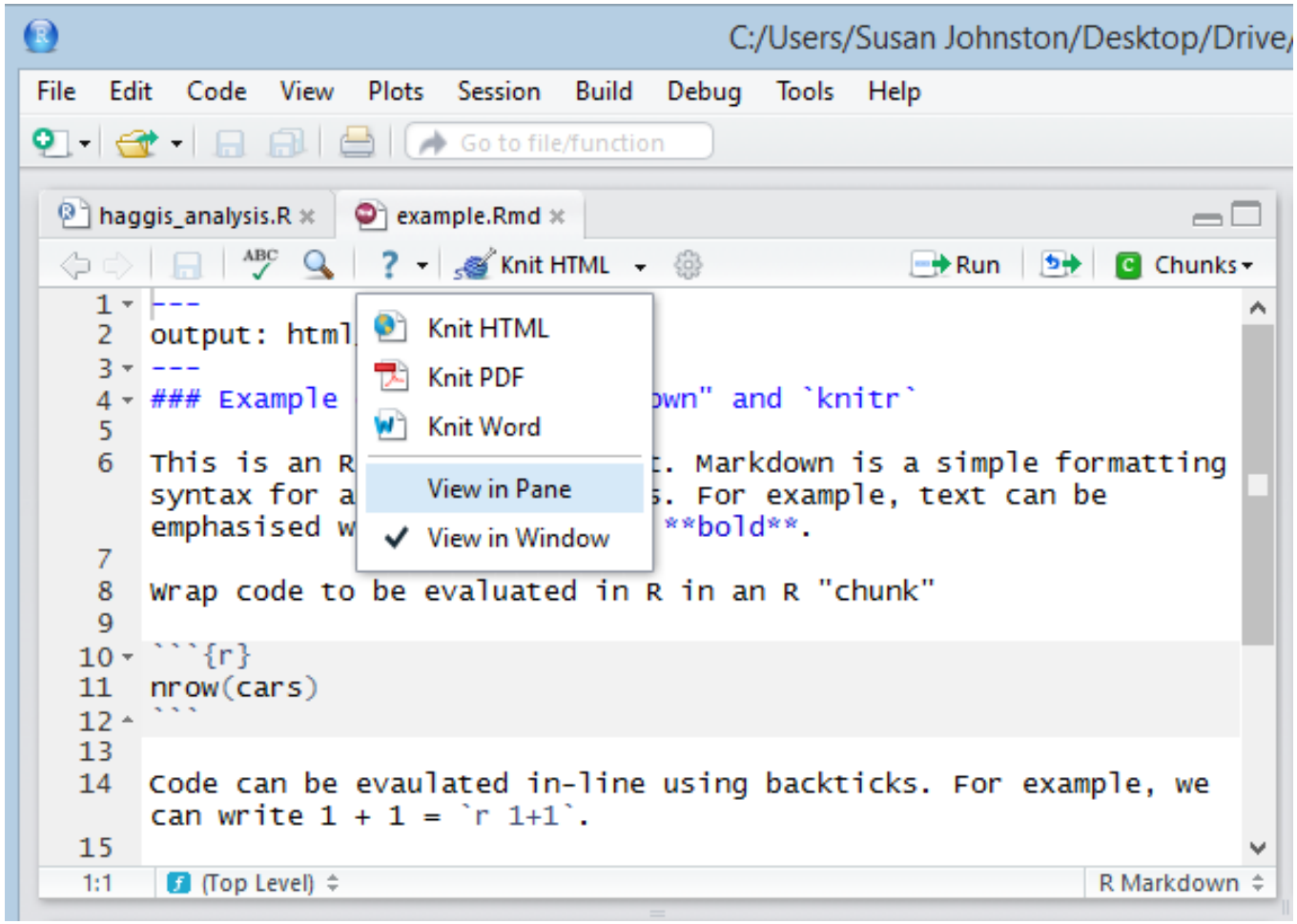
````
The slope of the model is `r coefficients(model1)[2]`
````

The slope of the model is 0.16557

4. Compile report as .html, .pdf or .doc

A Quick Start Guide

http://nicercode.github.io/guides/reports/

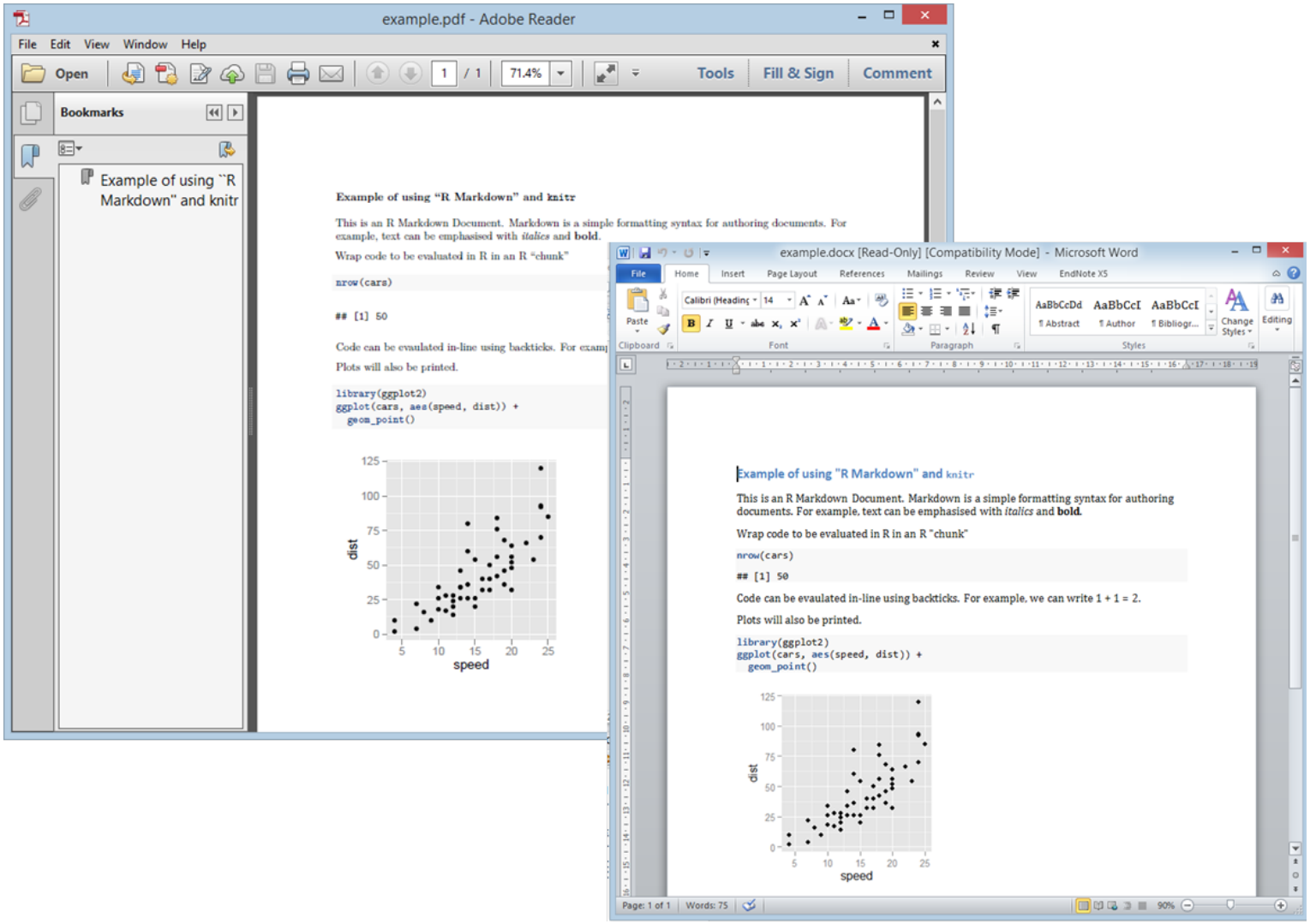


NB. PDF and Word docs require additional software.

http://rmarkdown.rstudio.com/?version=0.98.1103&mode=desktop

A Quick Start Guide

http://nicercode.github.io/guides/reports/



http://rmarkdown.rstudio.com/?version=0.98.1103&mode=desktop

Advanced Tips

- Control how chunks are reported and evaluated

````
```{r echo = F, warning = F, fig.width = 3}
model1 <- lm(speed ~ dist, data = cars)
plot(model1)
```
````

- `spin()`: compile .R files using #', #+ and #-

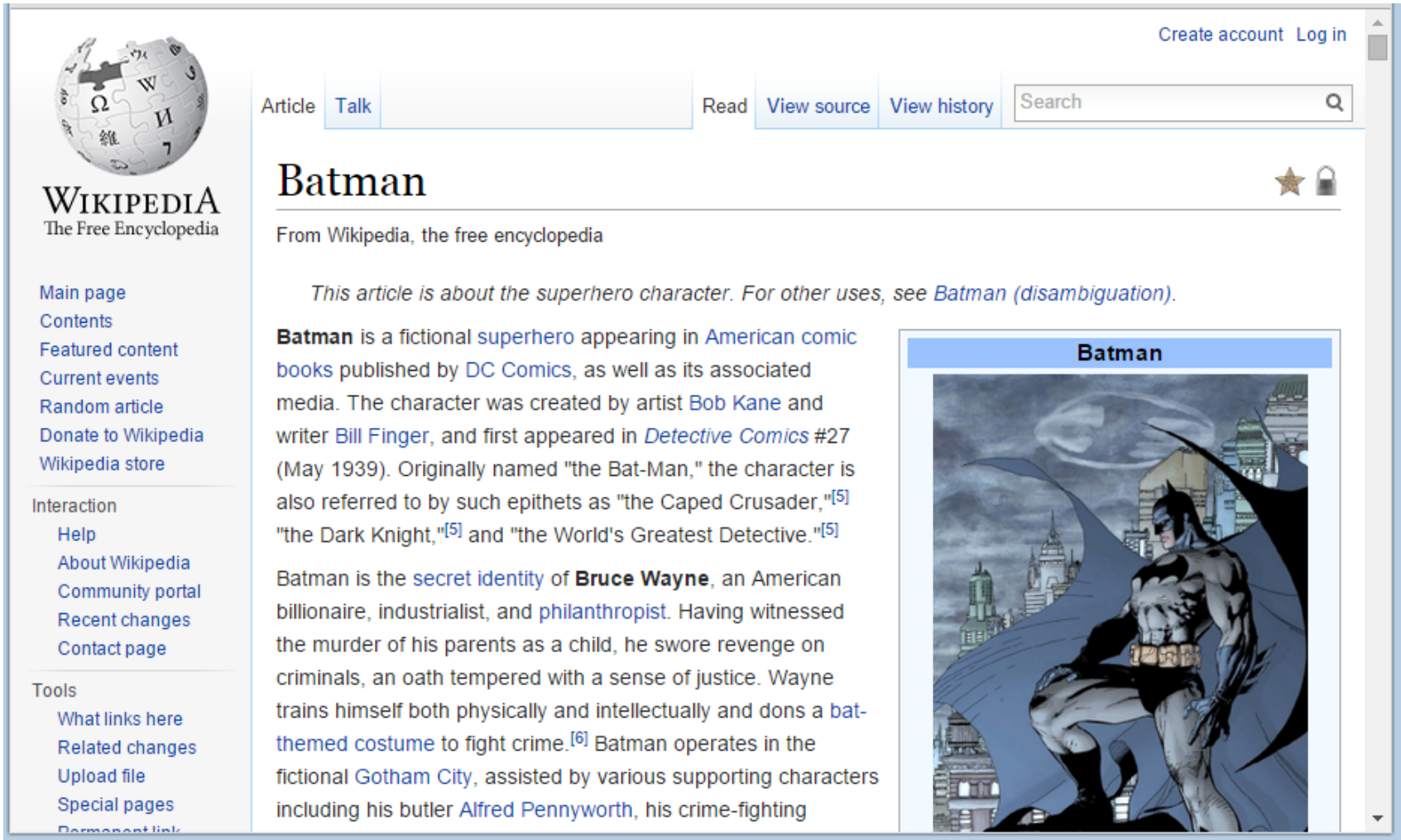
  http://deanattali.com/2015/03/24/knitrs-best-hidden-gem-spin/

- LaTeX documents, Presentations, Shiny, etc.

Reproducible Research in RStudio.

1. Creating a Portable Project (.Rproj)
2. Automate analyses - stop clicking and start typing.
3. Dynamic report writing with R Markdown and `knitr`
4. Version control using `git`

Version Control can revert a document to a previous version.



Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Talk Read View source View history Search

Batman

From Wikipedia, the free encyclopedia

This article is about the superhero character. For other uses, see Batman (disambiguation).

Batman is a fictional superhero appearing in American comic books published by DC Comics, as well as its associated media. The character was created by artist Bob Kane and writer Bill Finger, and first appeared in *Detective Comics* #27 (May 1939). Originally named "the Bat-Man," the character is also referred to by such epithets as "the Caped Crusader,"[5] "the Dark Knight,"[5] and "the World's Greatest Detective."[5]

Batman is the secret identity of **Bruce Wayne**, an American billionaire, industrialist, and philanthropist. Having witnessed the murder of his parents as a child, he swore revenge on criminals, an oath tempered with a sense of justice. Wayne trains himself both physically and intellectually and dons a bat-themed costume to fight crime.[6] Batman operates in the fictional Gotham City, assisted by various supporting characters including his butler Alfred Pennyworth, his crime-fighting

Batman

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages

Version Control can revert a document to a previous version.

Recent changes
Contact page

Tools
What links here
Related changes
Atom
Upload file
Special pages
Page information
Wikidata item

Languages

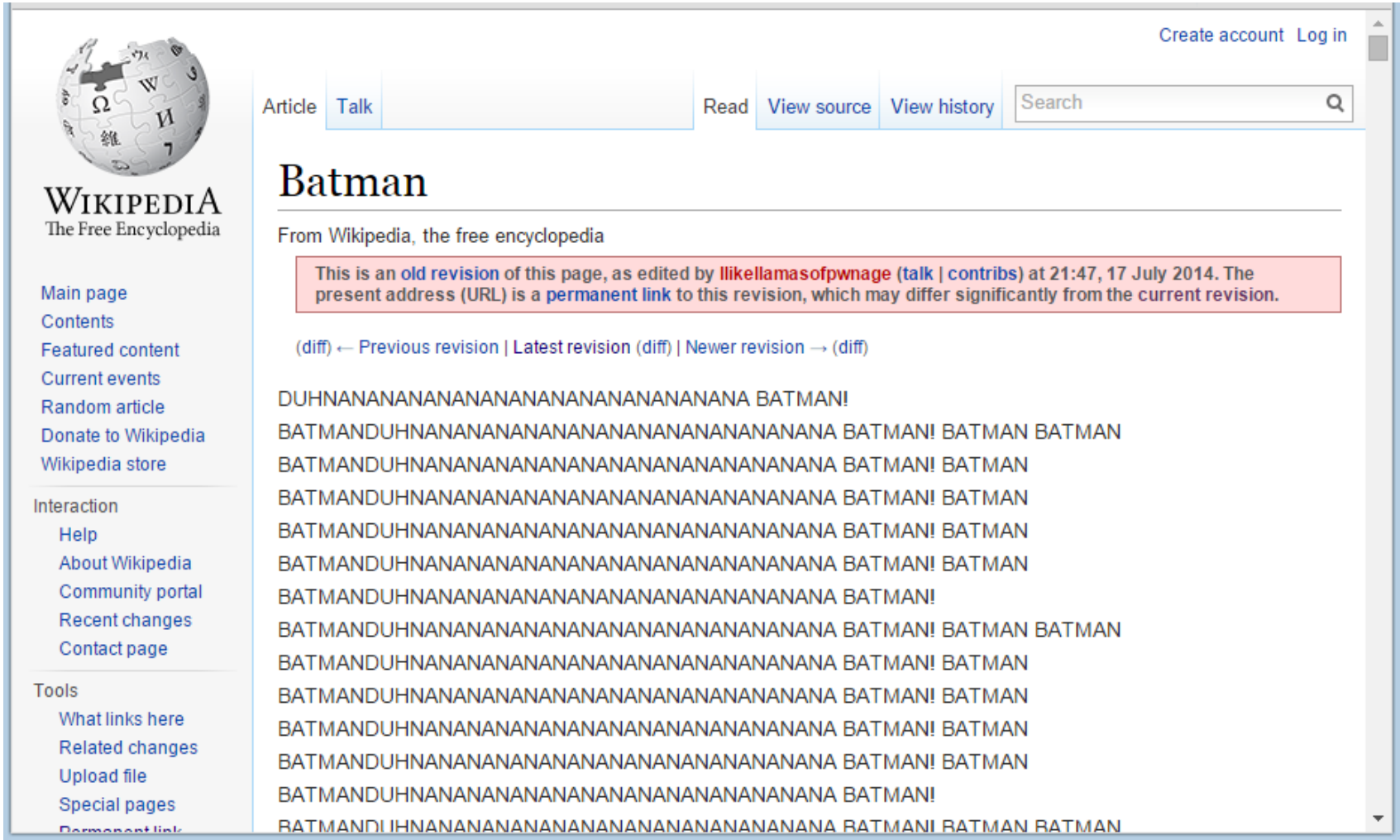
m = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

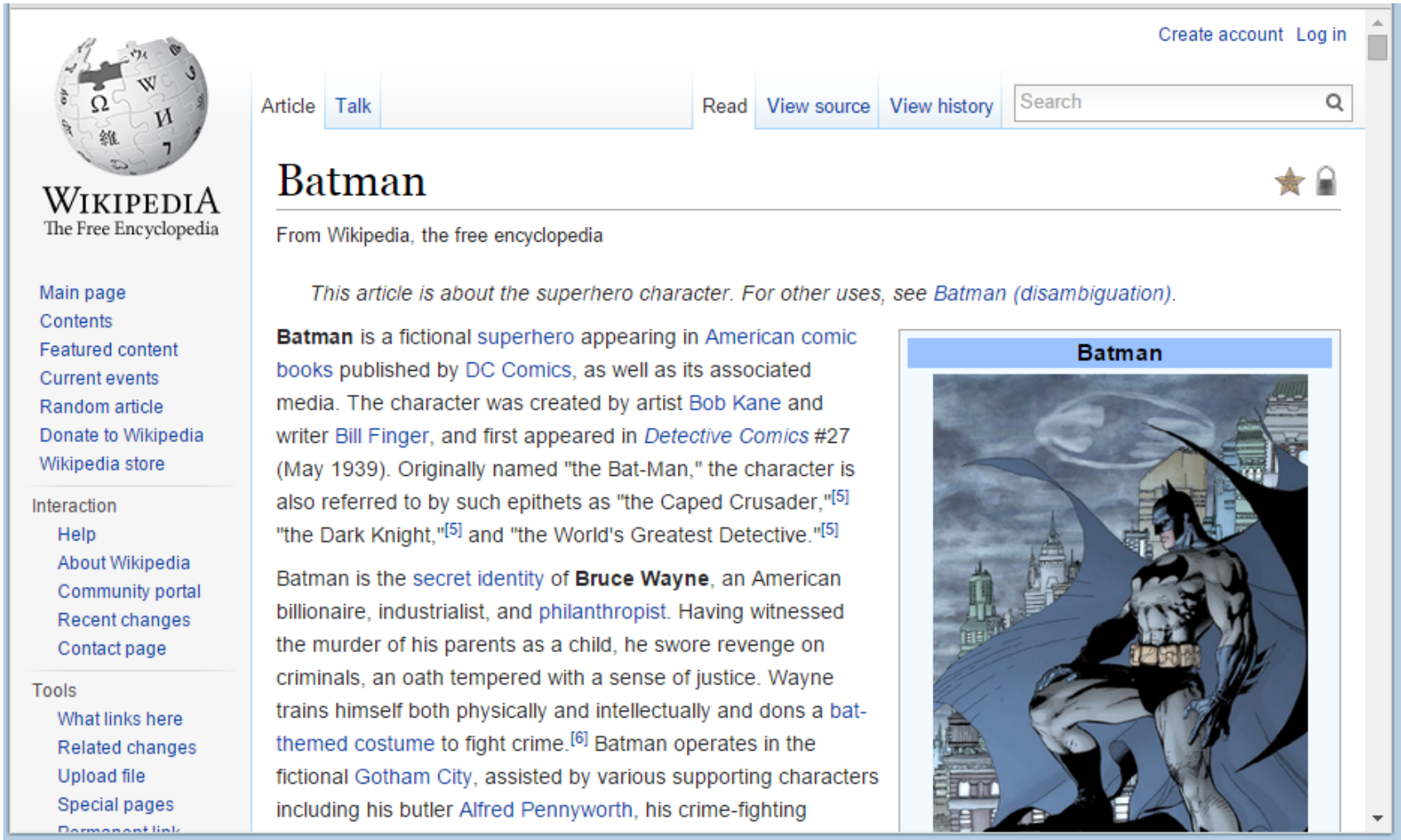
Compare selected revisions

- (cur | prev) 14:23, 15 April 2015 Kinfoll1993 (talk | contribs) **m** . . (118,704 bytes) (+10) . . *(→Personality)*
- (cur | prev) 14:20, 15 April 2015 Kinfoll1993 (talk | contribs) **m** . . (118,694 bytes) (+5) . . *(→Personality)*
- (cur | prev) 14:20, 15 April 2015 Kinfoll1993 (talk | contribs) **m** . . (118,689 bytes) (+1) . . *(→Personality)*
- (cur | prev) 14:19, 15 April 2015 Kinfoll1993 (talk | contribs) . . (118,688 bytes) **(+1,063)** . . *(→Personality: More info on Miller's Batman, original personality)*
- (cur | prev) 13:25, 15 April 2015 Kinfoll1993 (talk | contribs) **m** . . (117,625 bytes) (+34) . . *(Category:fictional antiheroes per citations at list of fictional antiheroes)*
- (cur | prev) 10:58, 15 April 2015 ProspectofArts (talk | contribs) . . (117,591 bytes) (+4)
- (cur | prev) 10:57, 15 April 2015 ProspectofArts (talk | contribs) . . (117,587 bytes) (+67)
- (cur | prev) 15:22, 9 April 2015 JosephSpiral (talk | contribs) . . (117,520 bytes) (+2)
- (cur | prev) 11:24, 7 April 2015 Brandmeister (talk | contribs) . . (117,518 bytes) **(-2,880)** . . *(more concise plot per tag)*
- (cur | prev) 23:42, 4 April 2015 Samhan Dobo (talk | contribs) . . (120,398 bytes) (+4) . . *(→Skills and abilities)*
- (cur | prev) 14:25, 3 April 2015 JosephSpiral (talk | contribs) . . (120,394 bytes) (0) . . *(→Romantic relationships)*
- (cur | prev) 14:24, 3 April 2015 JosephSpiral (talk | contribs) . . (120,394 bytes) (+21) . . *(→Romantic*

Version Control can revert a document to a previous version.



Version Control can revert a document to a previous version.



Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Talk Read View source View history Search

Batman

From Wikipedia, the free encyclopedia

This article is about the superhero character. For other uses, see Batman (disambiguation).

Batman is a fictional superhero appearing in American comic books published by DC Comics, as well as its associated media. The character was created by artist Bob Kane and writer Bill Finger, and first appeared in *Detective Comics* #27 (May 1939). Originally named "the Bat-Man," the character is also referred to by such epithets as "the Caped Crusader,"[5] "the Dark Knight,"[5] and "the World's Greatest Detective."[5]

Batman is the secret identity of **Bruce Wayne**, an American billionaire, industrialist, and philanthropist. Having witnessed the murder of his parents as a child, he swore revenge on criminals, an oath tempered with a sense of justice. Wayne trains himself both physically and intellectually and dons a bat-themed costume to fight crime.[6] Batman operates in the fictional Gotham City, assisted by various supporting characters including his butler Alfred Pennyworth, his crime-fighting

Batman

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages

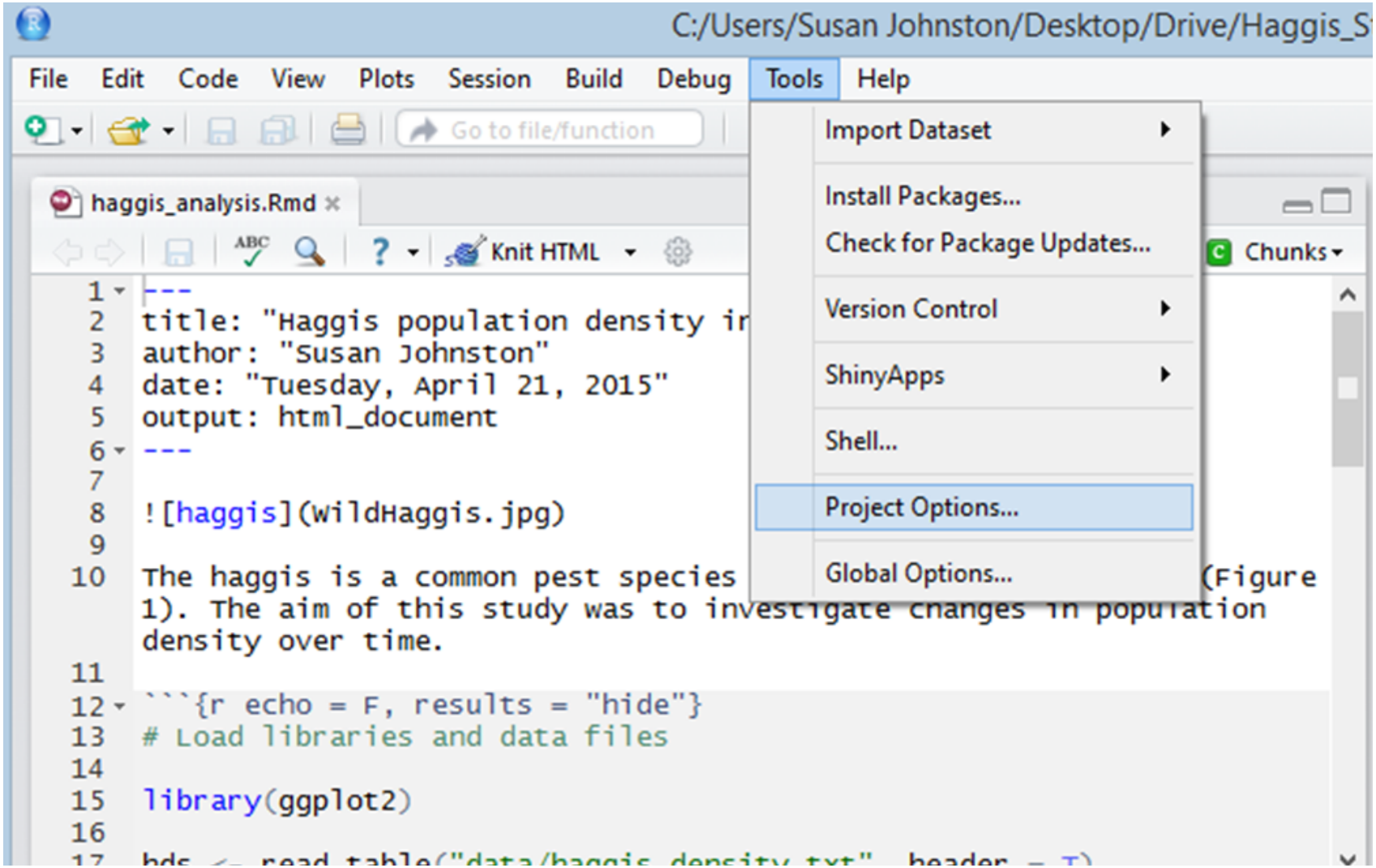
Version Control Using git.

https://support.rstudio.com/hc/en-us/articles/200532077-Version-Control-with-Git-and-SVN

Git can be installed on all platforms, and can be used to implement version control within an R Studio Project.

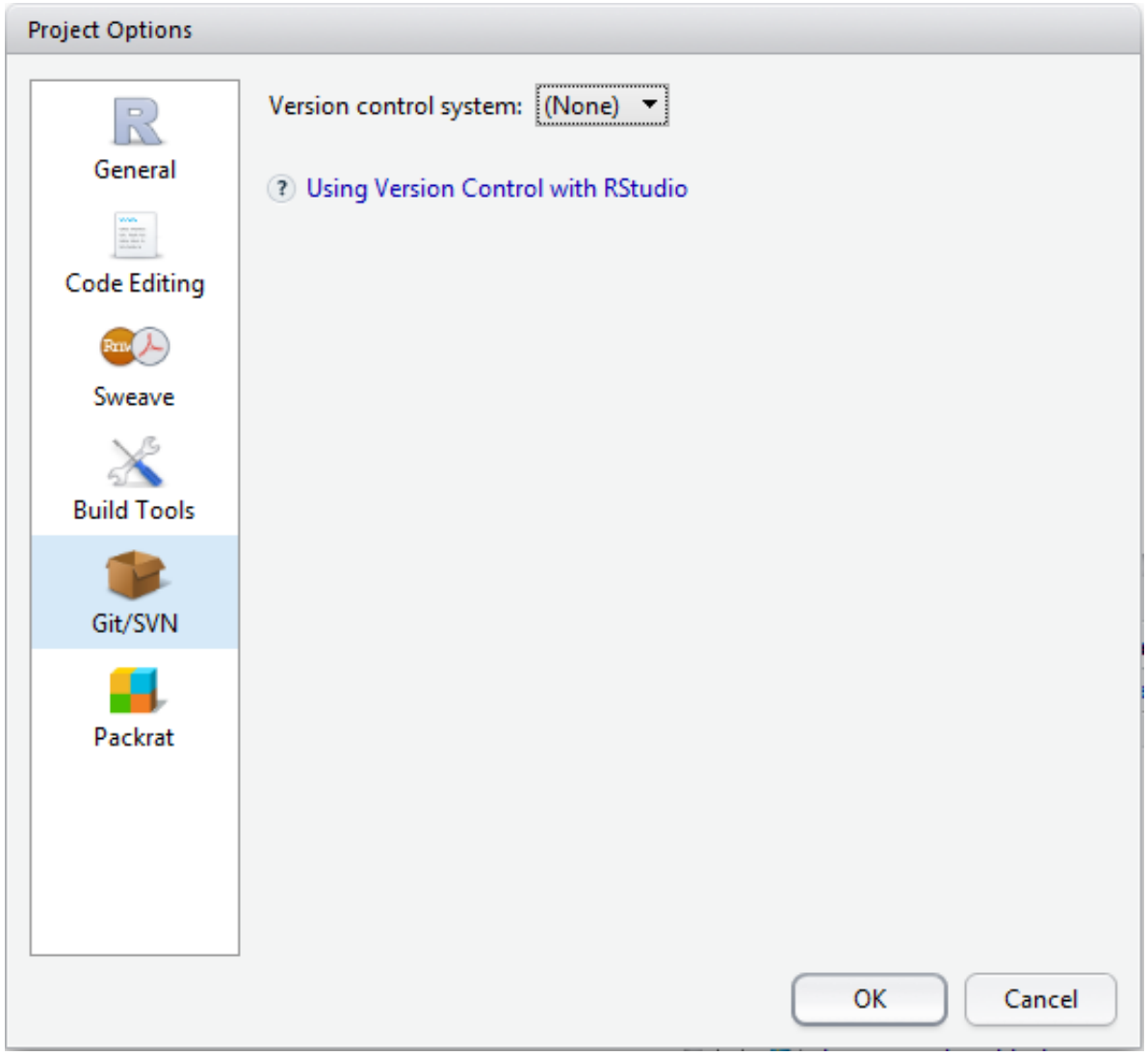
http://git-scm.com/downloads

Version Control in R Studio



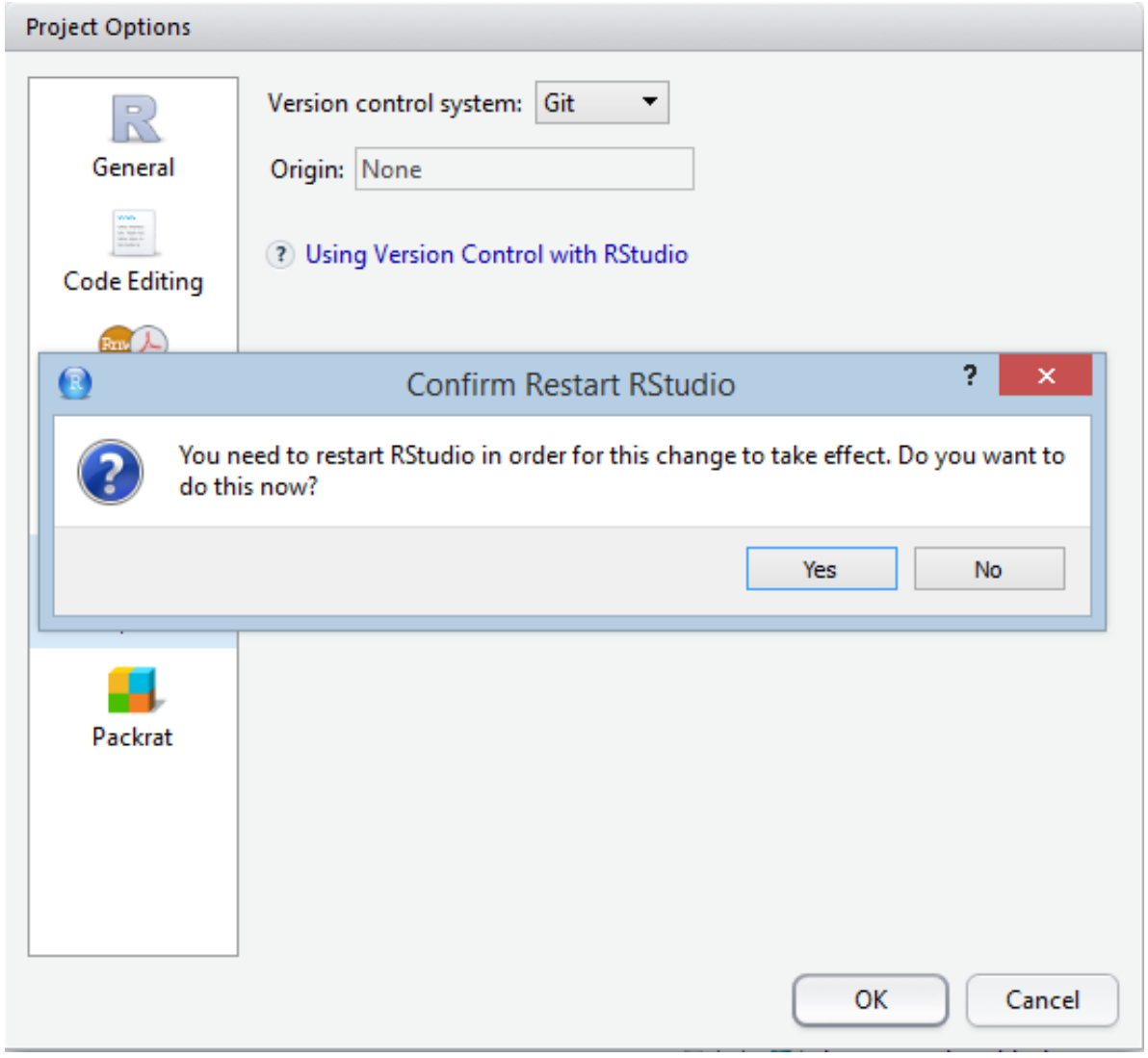
Tools >Project Options allows setup of `git` version control.

Version Control in R Studio



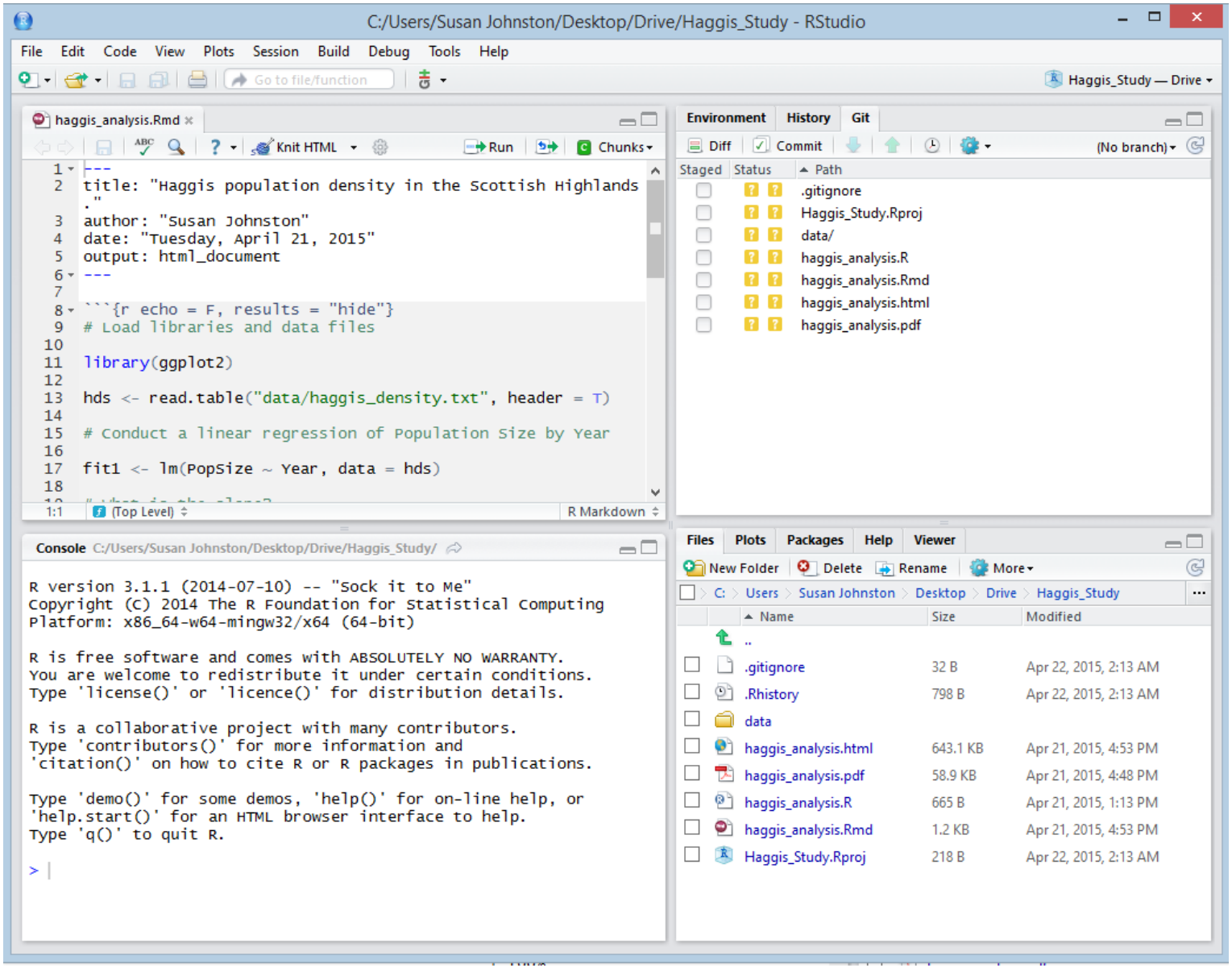
Select `git` as a version control system

Version Control in R Studio



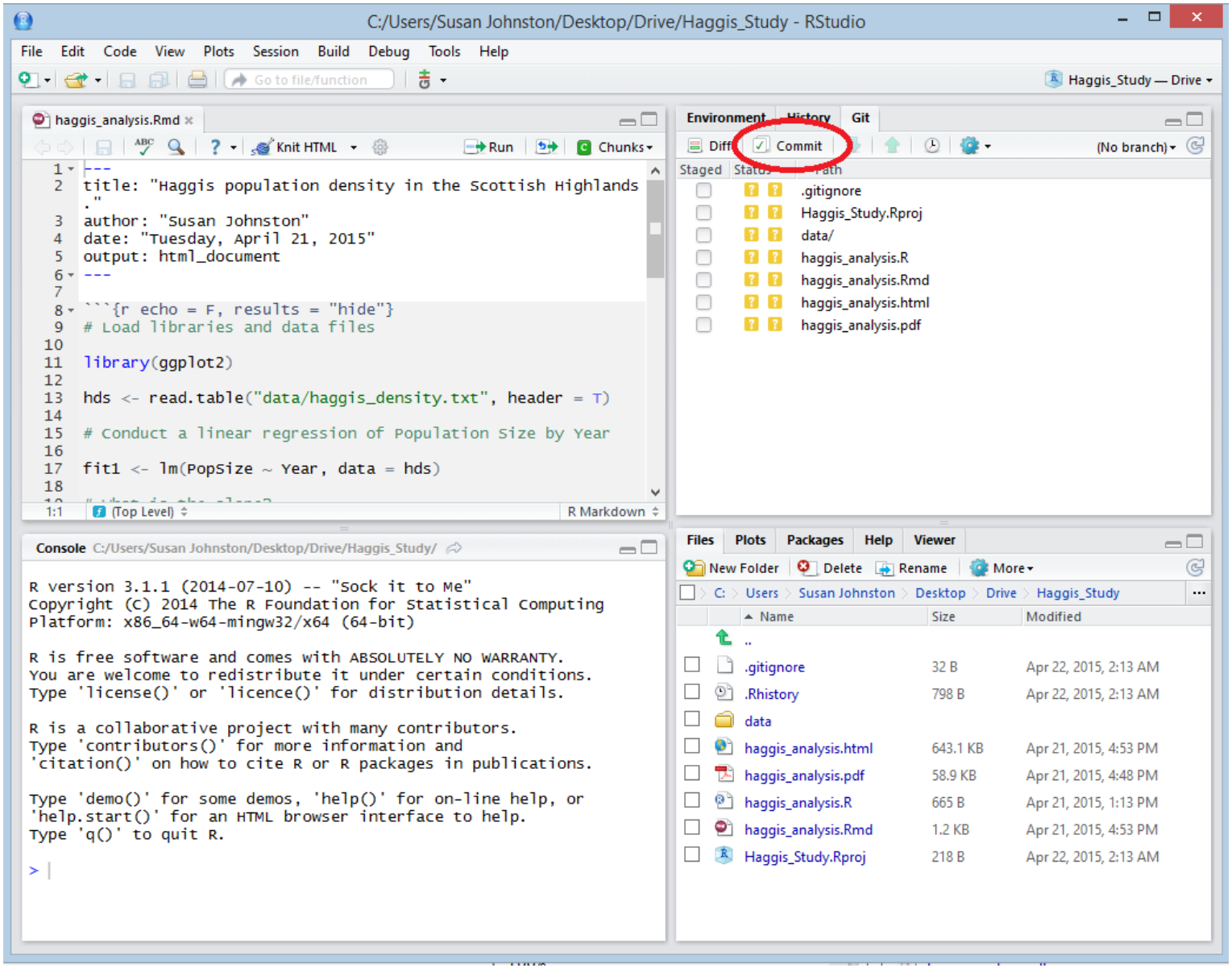
Select `git` as a version control system

Version Control in R Studio



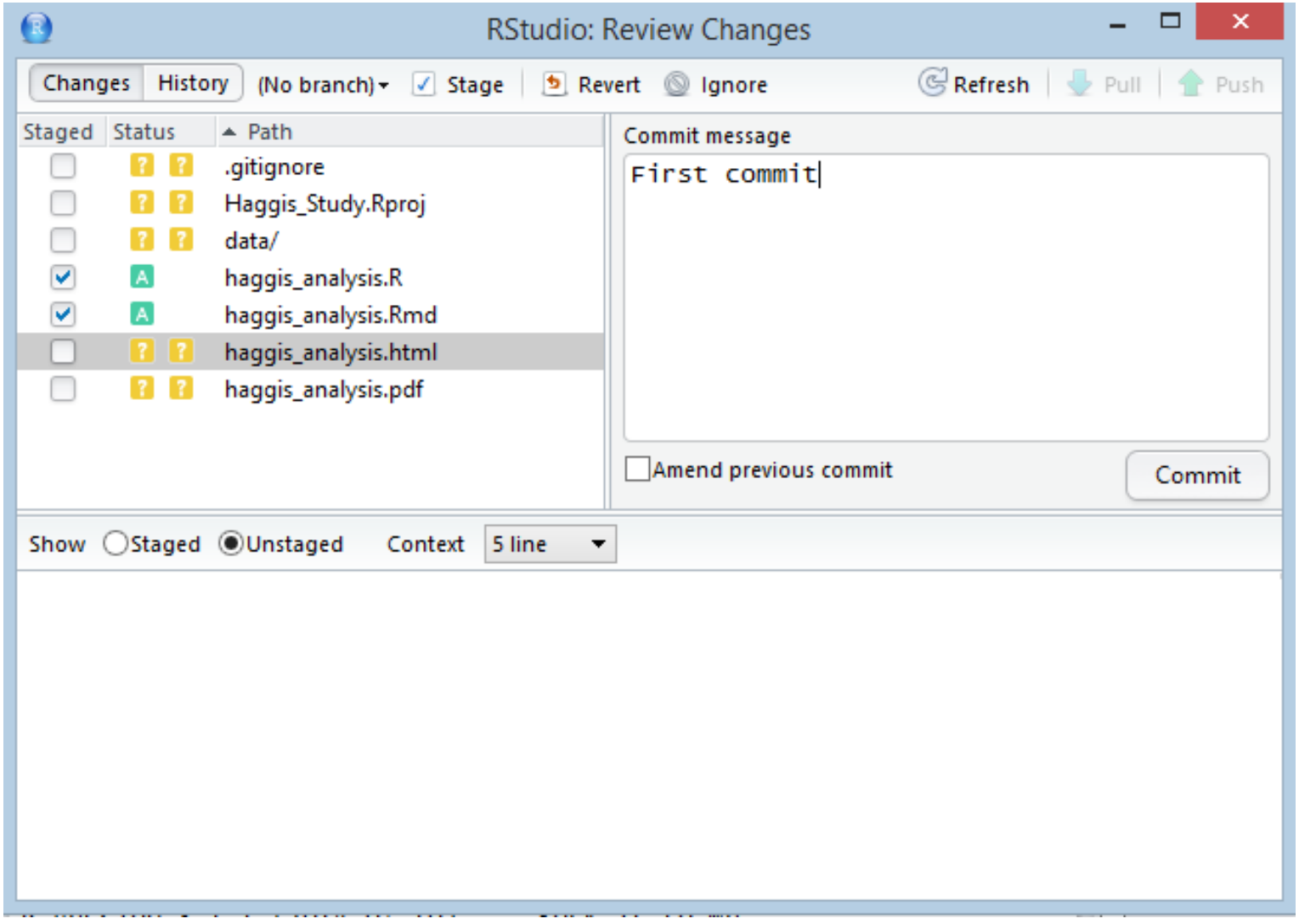
`git` information will appear in the top-right frame.

Version Control in R Studio



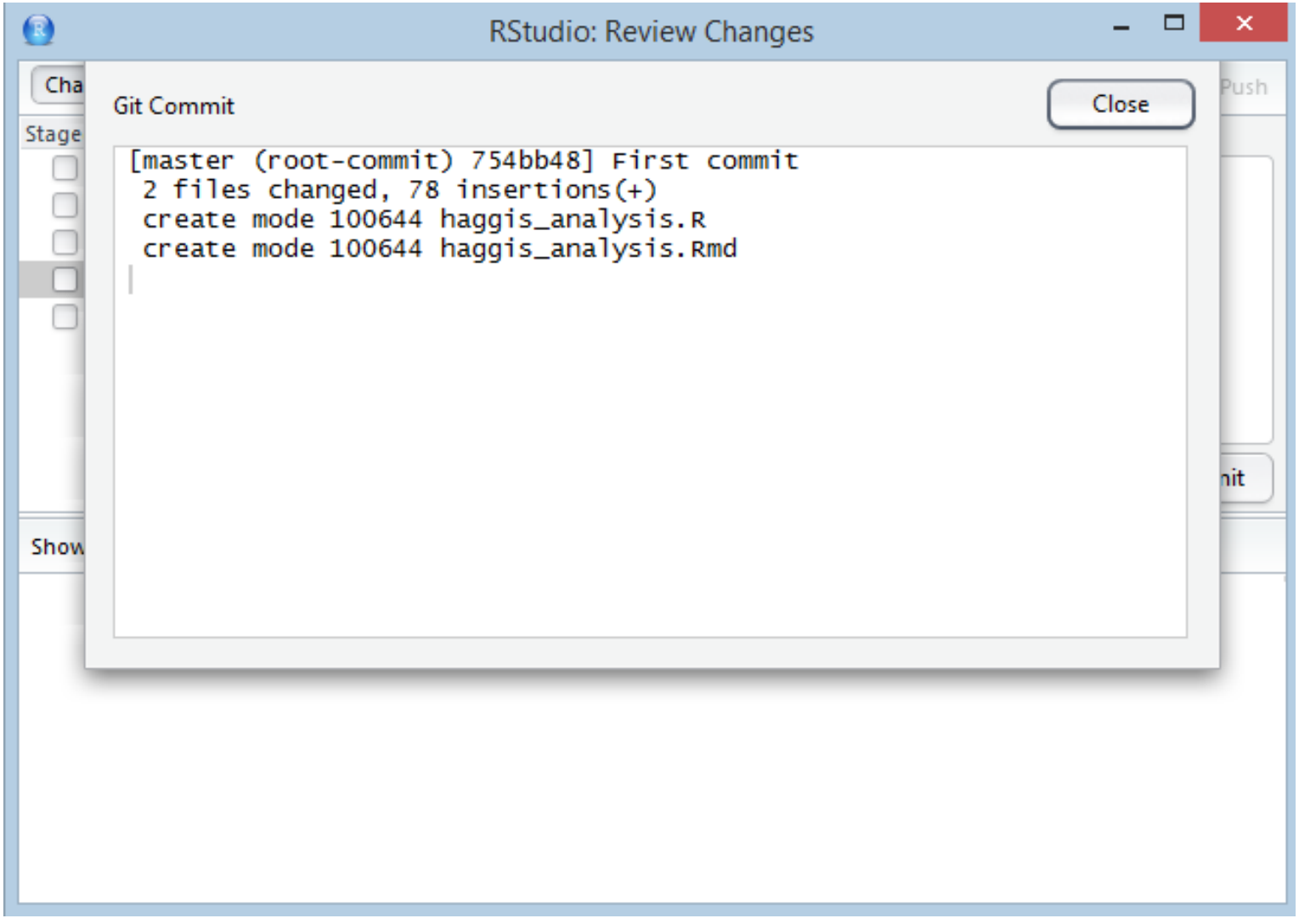
`git` information will appear in the top-right frame.

Version Control in R Studio



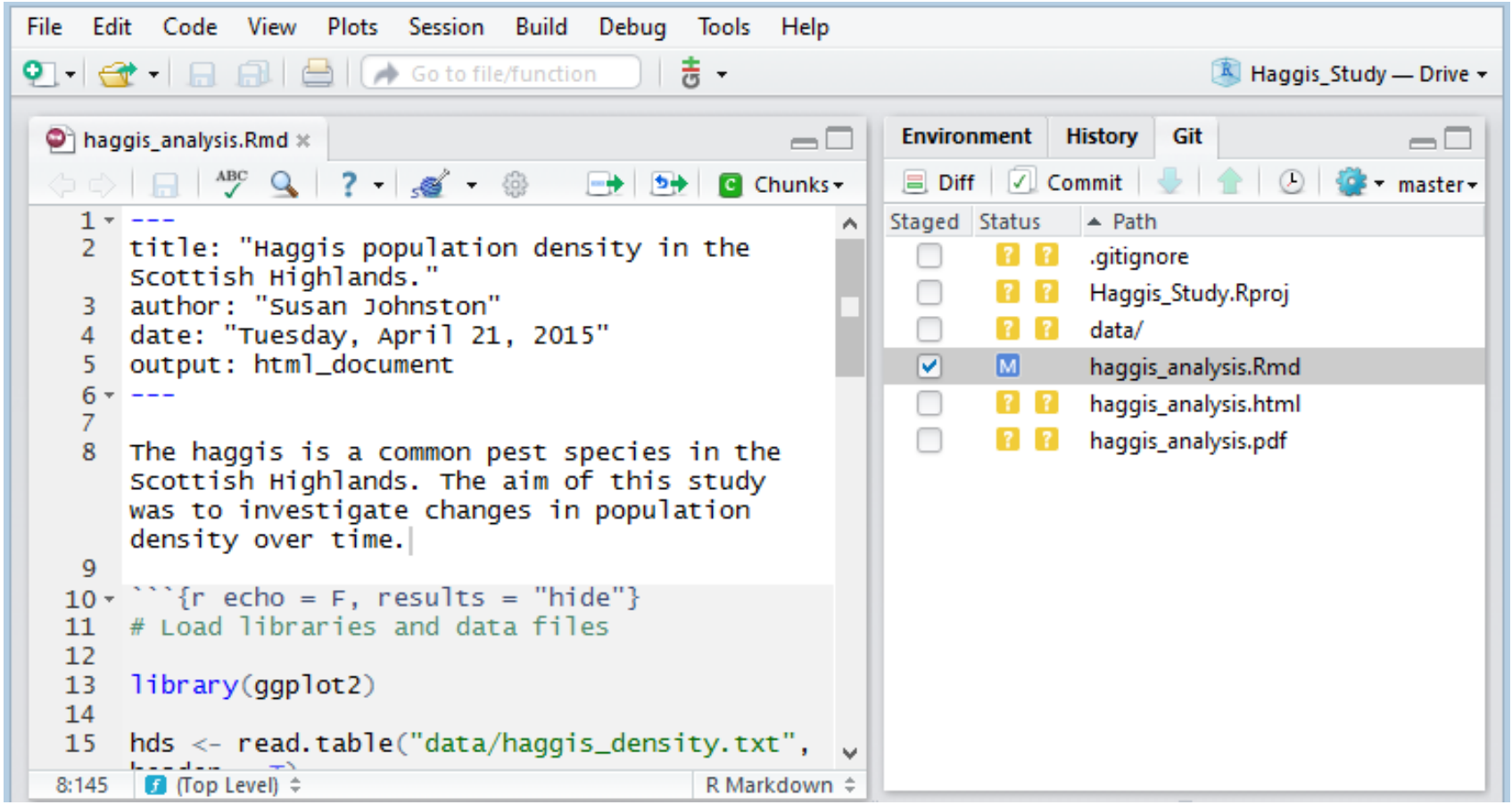
Select files to version control, write a meaningful commit message >Commit

Version Control in R Studio



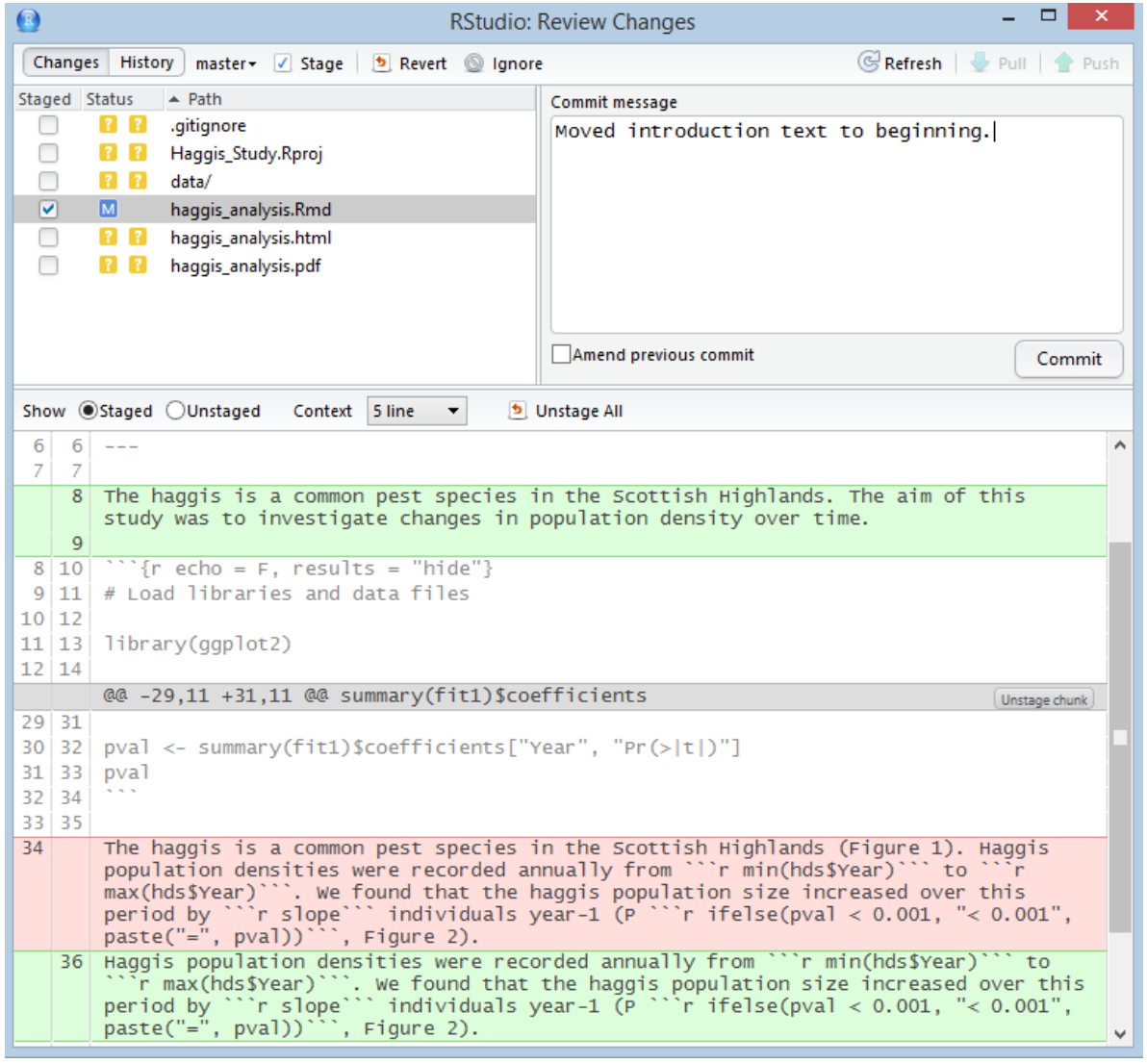
Select files to version control, write a meaningful commit message >Commit

Version Control in R Studio



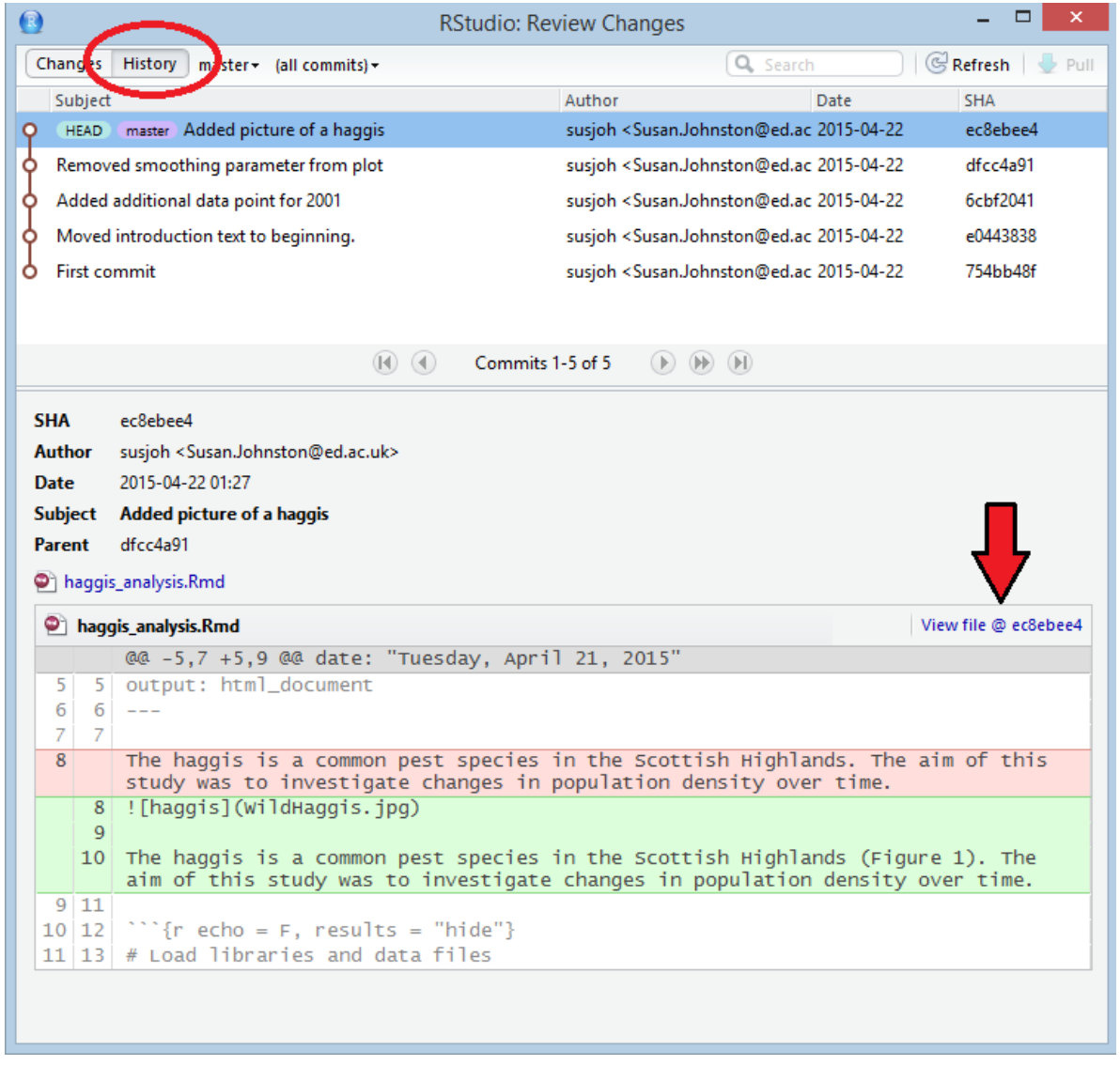
After modifying the file, repeat the process.

Version Control in R Studio



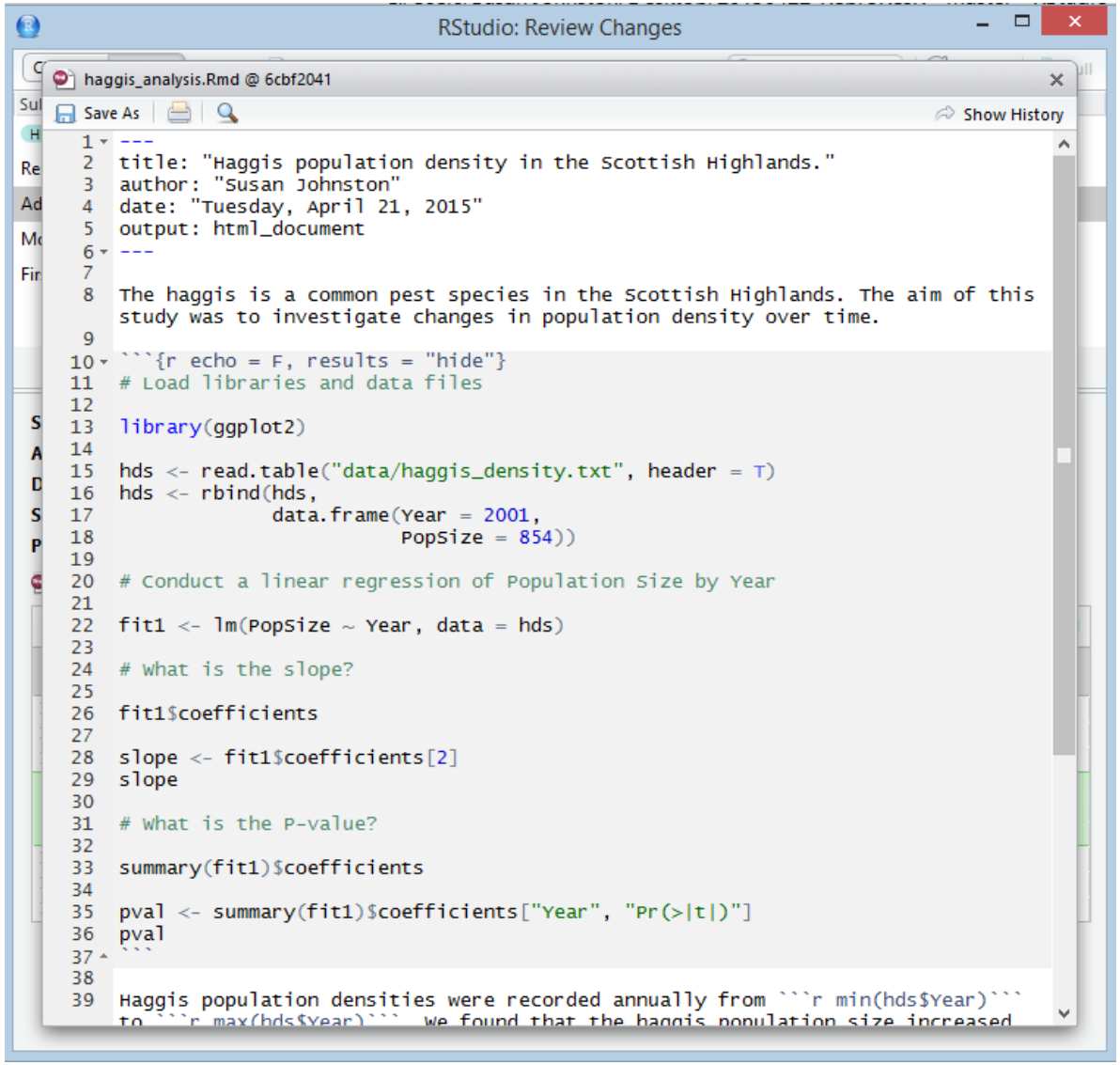
After modifying the file, repeat the process.

Version Control in R Studio



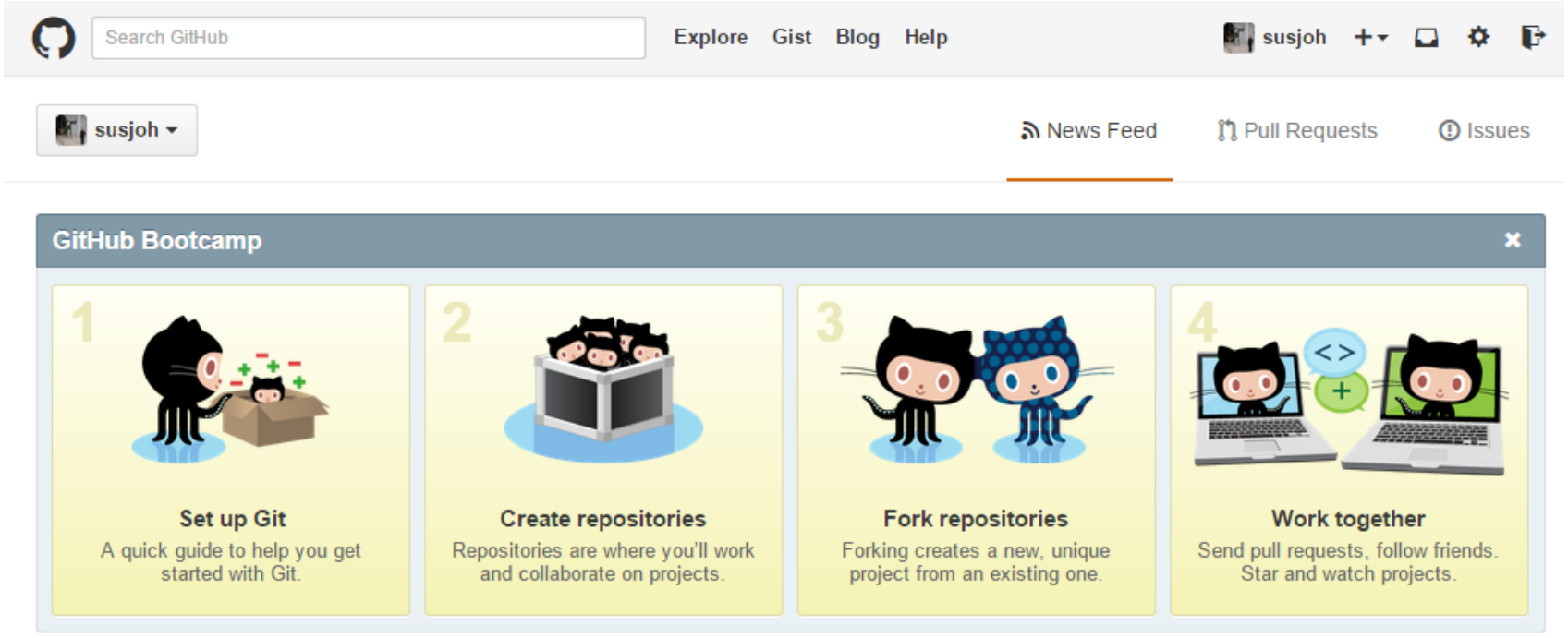
Previous versions can be viewed and restored from the History tab.

Version Control in R Studio



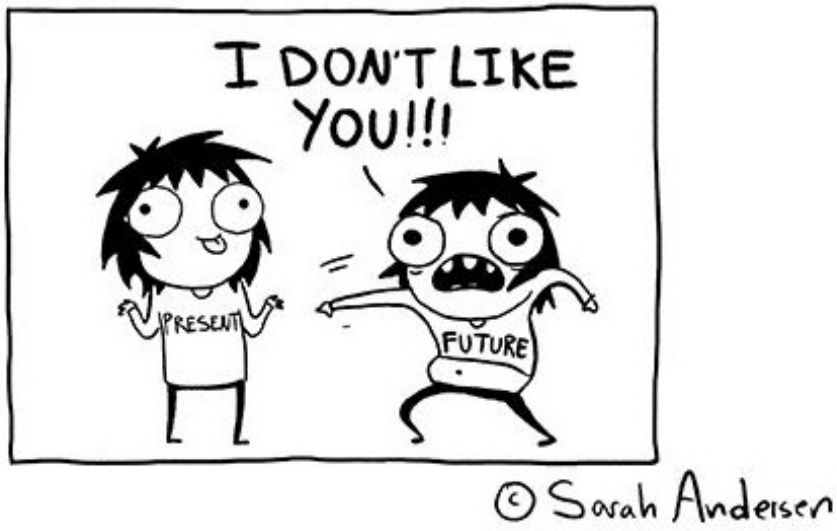
Previous versions can be viewed and restored from the History tab.

Advanced Steps: Github



- Forking projects
- All scripts are backed up online
- Facilitates collaboration and working on different computers

Take home messages



- Manage projects reproducibly: The first researcher who will need to reproduce the results is likely to be **YOU**.
- Time invested in learning to code pays off - **do it**.
- Supervisors should be patient and encourage students to code.

Online Resources

- RStudio: Idiot-proof guides and cheat sheets
  http://www.rstudio.com/
- Nice R Code: How-tos and advice on good coding practice
  http://nicercode.github.io/guide.html
- Ten Simple Rules for Reproducible Computational Research
  http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285
- Yihui Xie's blog (knitr) http://yihui.name/en/categories/
- R Bloggers: http://www.r-bloggers.com/
- StackOverflow questions on R and knitr
  http://stackoverflow.com/questions/tagged/r+knitr