Achieving reproducibility

Jeff Leek

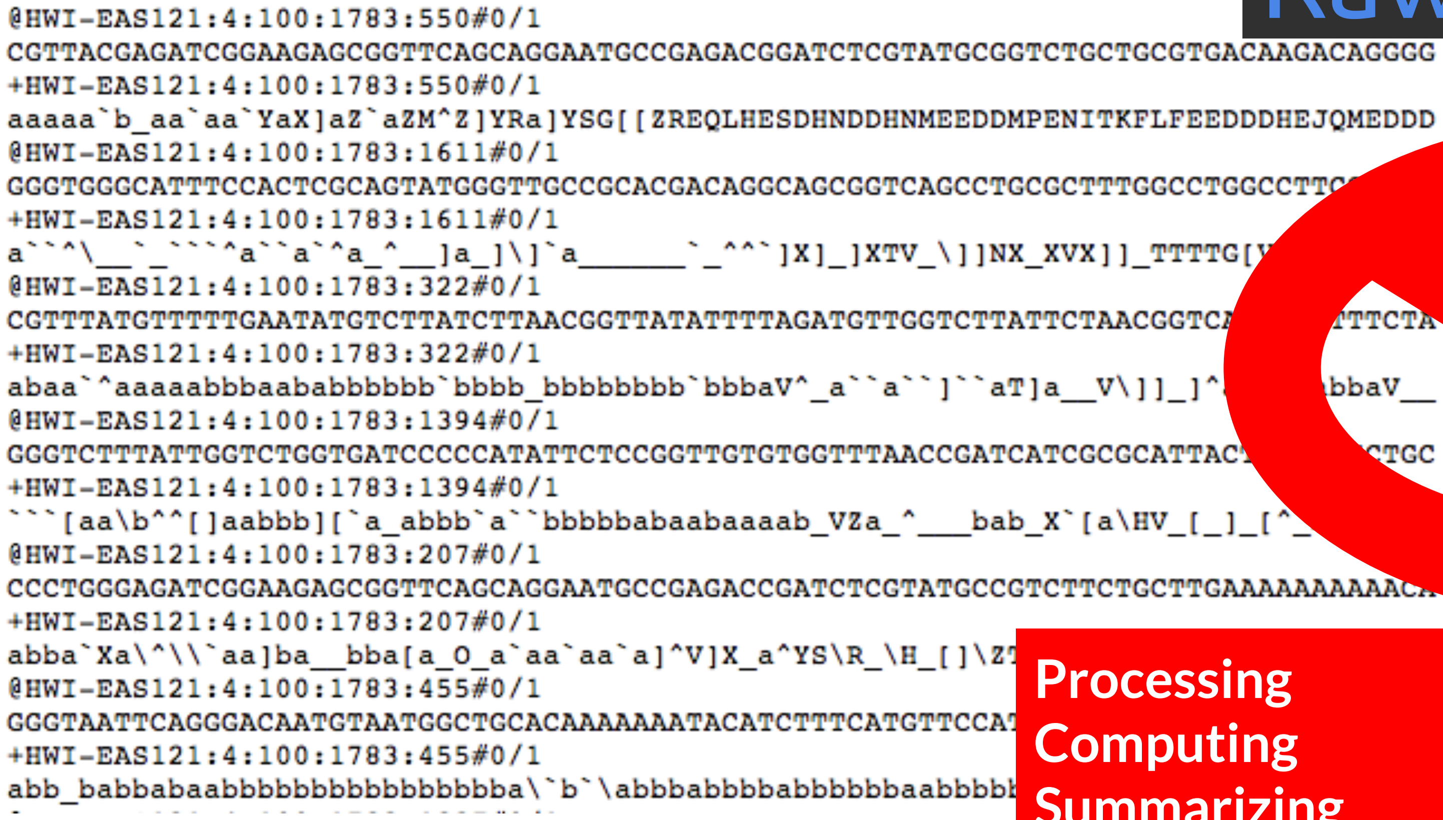
@jtleek

www.jtleek.com

A data sharing plan

1. The raw data.
2. A tidy data set
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 -> 2,3

Raw data

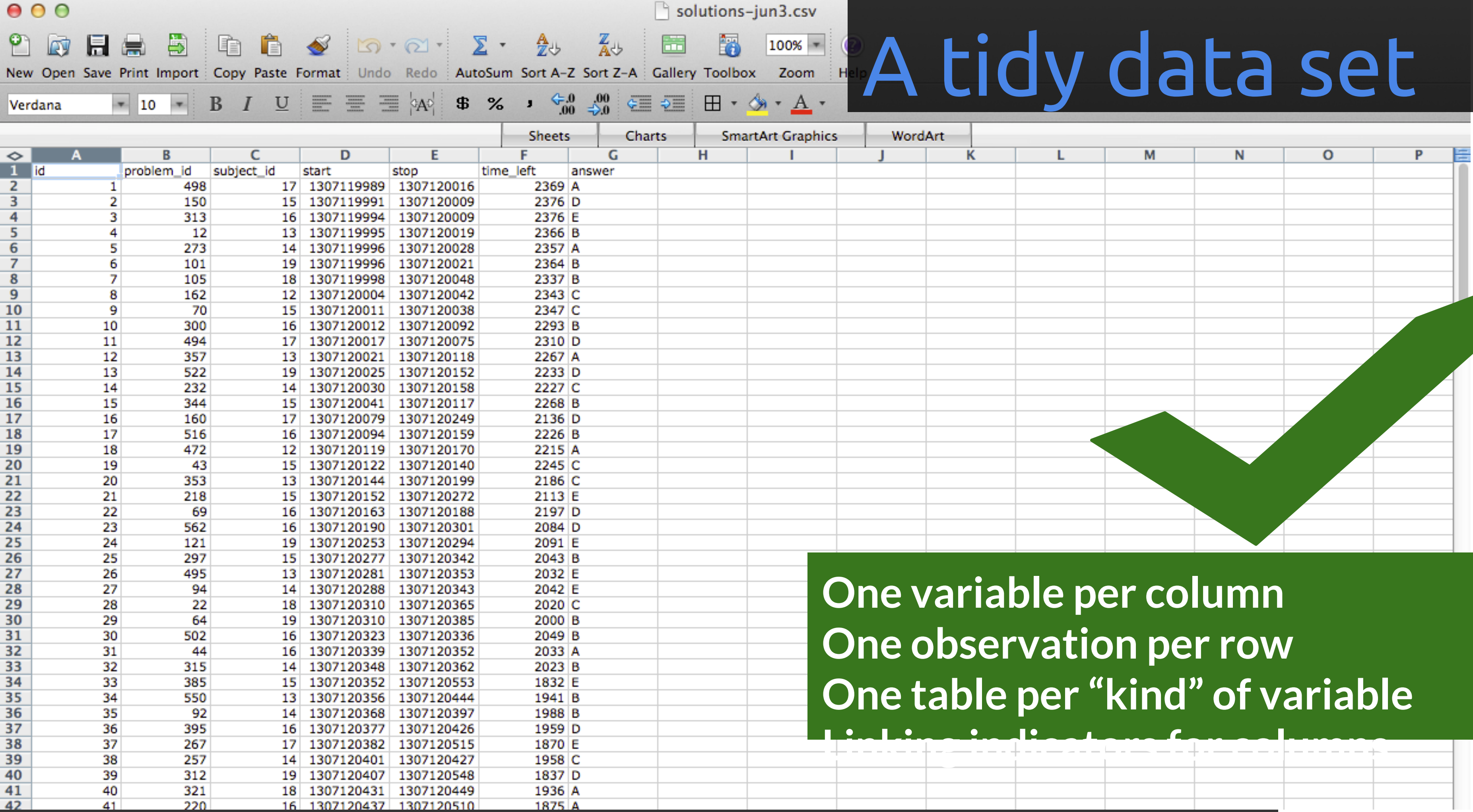


~~~
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTC
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a_______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[V
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCA     TTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^     bbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTAC     CTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\Z
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCA
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbb
~~~

Processing
Computing
Summarizing
Deleting

Reference: http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt

A tidy data set

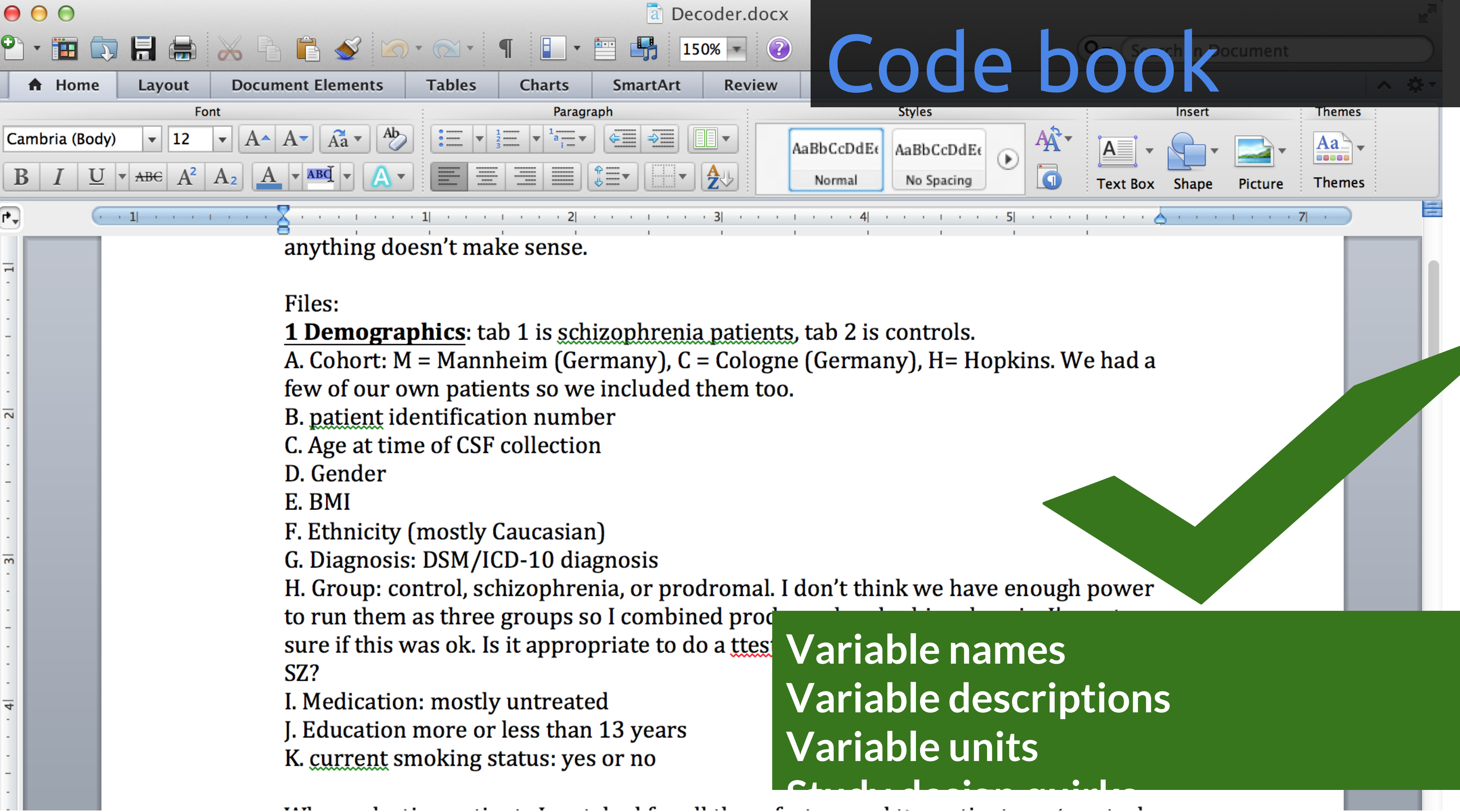


id	problem_id	subject_id	start	stop	time_left	answer
1	498	17	1307119989	1307120016	2369	A
2	150	15	1307119991	1307120009	2376	D
3	313	16	1307119994	1307120009	2376	E
4	12	13	1307119995	1307120019	2366	B
5	273	14	1307119996	1307120028	2357	A
6	101	19	1307119996	1307120021	2364	B
7	105	18	1307119998	1307120048	2337	B
8	162	12	1307120004	1307120042	2343	C
9	70	15	1307120011	1307120038	2347	C
10	300	16	1307120012	1307120092	2293	B
11	494	17	1307120017	1307120075	2310	D
12	357	13	1307120021	1307120118	2267	A
13	522	19	1307120025	1307120152	2233	D
14	232	14	1307120030	1307120158	2227	C
15	344	15	1307120041	1307120117	2268	B
16	160	17	1307120079	1307120249	2136	D
17	516	16	1307120094	1307120159	2226	B
18	472	12	1307120119	1307120170	2215	A
19	43	15	1307120122	1307120140	2245	C
20	353	13	1307120144	1307120199	2186	C
21	218	15	1307120152	1307120272	2113	E
22	69	16	1307120163	1307120188	2197	D
23	562	16	1307120190	1307120301	2084	D
24	121	19	1307120253	1307120294	2091	E
25	297	15	1307120277	1307120342	2043	B
26	495	13	1307120281	1307120353	2032	E
27	94	14	1307120288	1307120343	2042	E
28	22	18	1307120310	1307120365	2020	C
29	64	19	1307120310	1307120385	2000	B
30	502	16	1307120323	1307120336	2049	B
31	44	16	1307120339	1307120352	2033	A
32	315	14	1307120348	1307120362	2023	B
33	385	15	1307120352	1307120553	1832	E
34	550	13	1307120356	1307120444	1941	B
35	92	14	1307120368	1307120397	1988	B
36	395	16	1307120377	1307120426	1959	D
37	267	17	1307120382	1307120515	1870	E
38	257	14	1307120401	1307120427	1958	C
39	312	19	1307120407	1307120548	1837	D
40	321	18	1307120431	1307120449	1936	A
41	220	16	1307120437	1307120510	1875	A

One variable per column
One observation per row
One table per “kind” of variable
Linking indicators for columns

Reference: http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt

Code book



anything doesn’t make sense.

Files:

1 Demographics: tab 1 is schizophrenia patients, tab 2 is controls.

A. Cohort: M = Mannheim (Germany), C = Cologne (Germany), H= Hopkins. We had a few of our own patients so we included them too.

B. patient identification number

C. Age at time of CSF collection

D. Gender

E. BMI

F. Ethnicity (mostly Caucasian)

G. Diagnosis: DSM/ICD-10 diagnosis

H. Group: control, schizophrenia, or prodromal. I don’t think we have enough power to run them as three groups so I combined prod

sure if this was ok. Is it appropriate to do a ttest

SZ?

I. Medication: mostly untreated

J. Education more or less than 13 years

K. current smoking status: yes or no

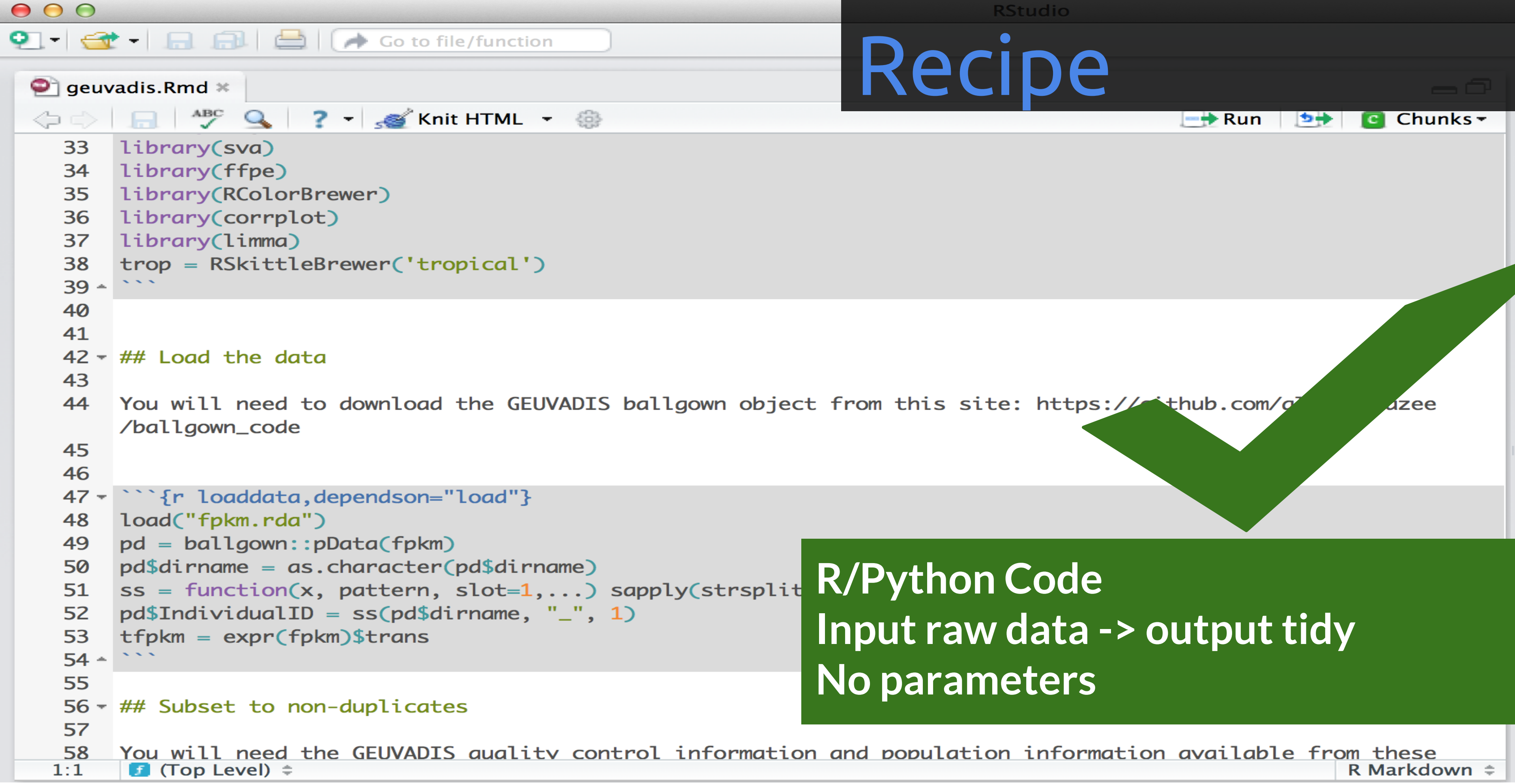
Variable names

Variable descriptions

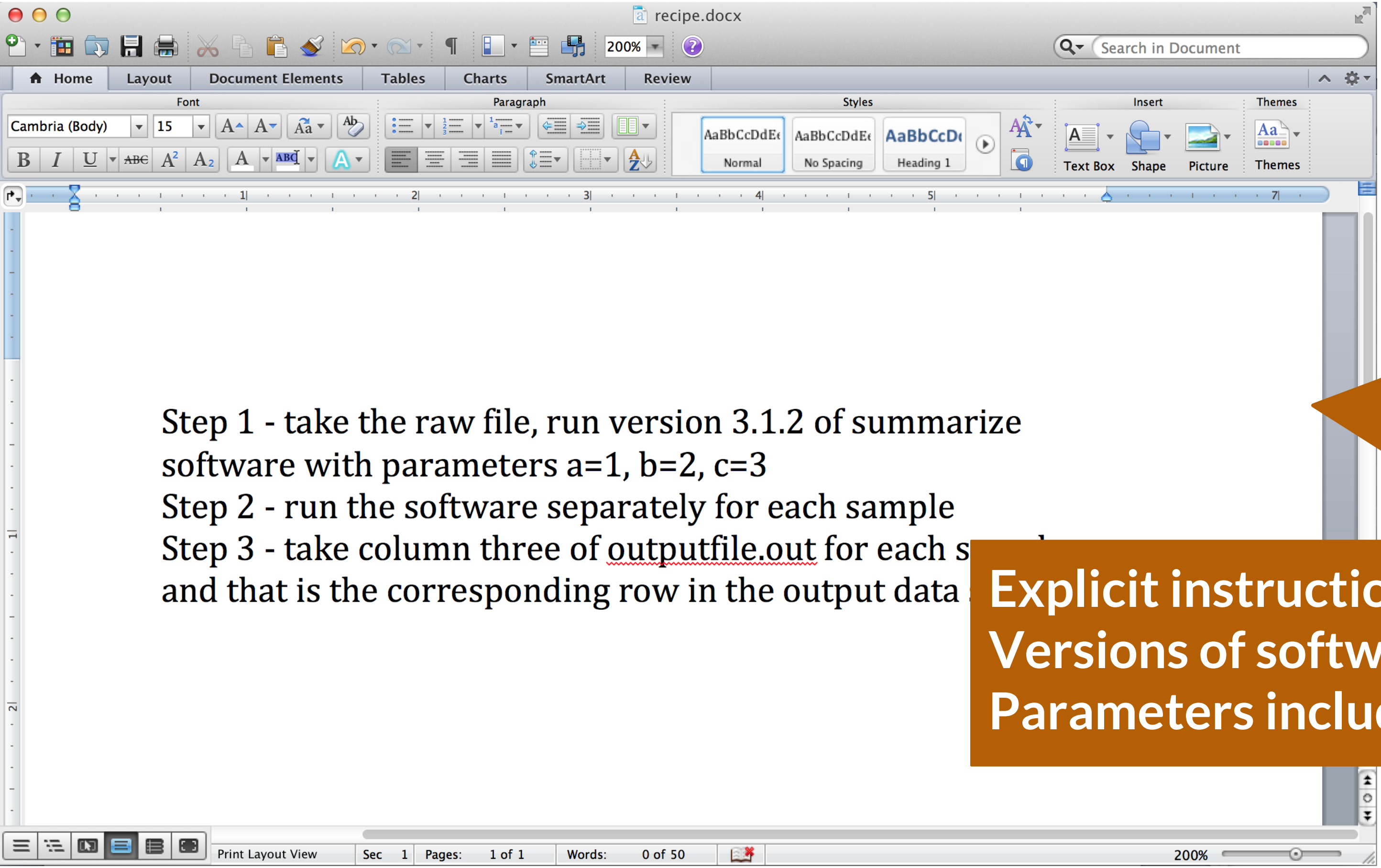
Variable units

Study design quirks

Recipe

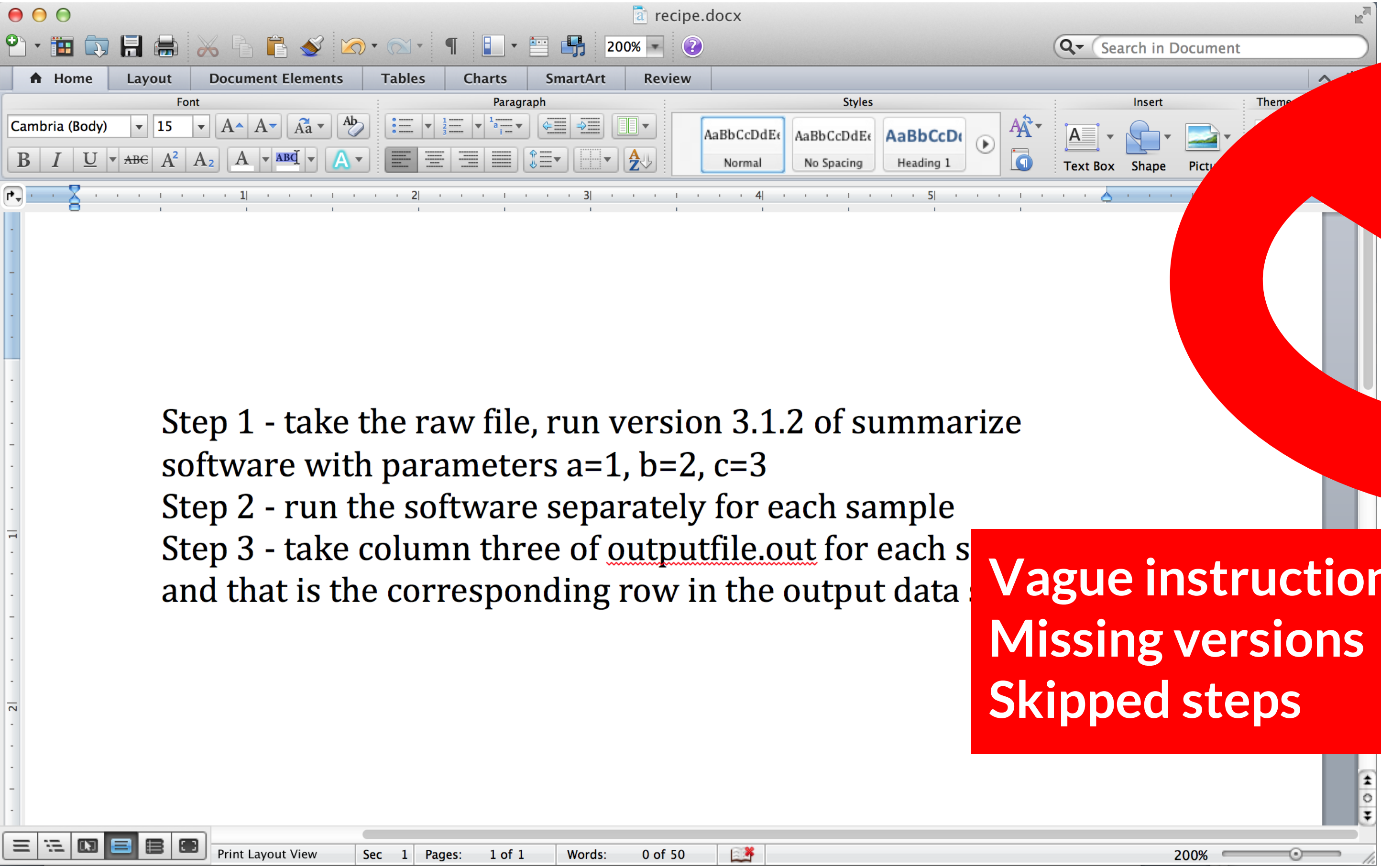


```
library(sva)
library(ffpe)
library(RColorBrewer)
library(corrplot)
library(limma)
trop = RSkittleBrewer('tropical')
```
```

Load the data

You will need to download the GEUVADIS ballgown object from this site: https://   ithub.com/a   azee /ballgown_code

```
```{r loaddata,dependson="load"}
load("fpkm.rda")
pd = ballgown::pData(fpkm)
pd$dirname = as.character(pd$dirname)
ss = function(x, pattern, slot=1,...) sapply(strsplit
pd$IndividualID = ss(pd$dirname, "_", 1)
tfpkm = expr(fpkm)$trans
```
```

Subset to non-duplicates

You will need the GEUVADIS quality control information and population information available from these

R/Python Code
Input raw data -> output tidy
No parameters
```



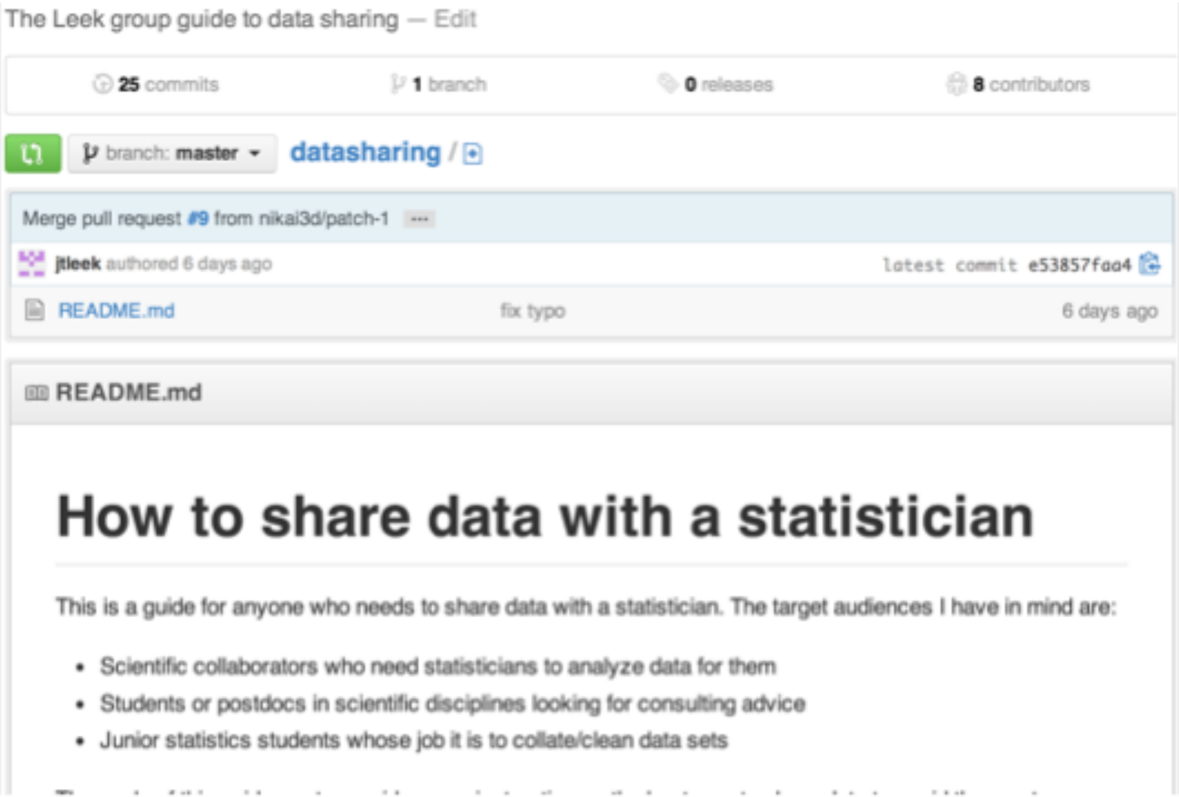
Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each s
and that is the corresponding row in the output data

Explicit instructions
Versions of software
Parameters included



Step 1 - take the raw file, run version 3.1.2 of summarize
software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each s
and that is the corresponding row in the output data

Vague instructions
Missing versions
Skipped steps



The Leek group guide to data sharing — Edit

25 commits | 1 branch | 0 releases | 8 contributors

branch: master | datasharing /

Merge pull request #9 from nikai3d/patch-1 …

jtleek authored 6 days ago — latest commit e53857faa4

README.md — fix typo — 6 days ago

README.md

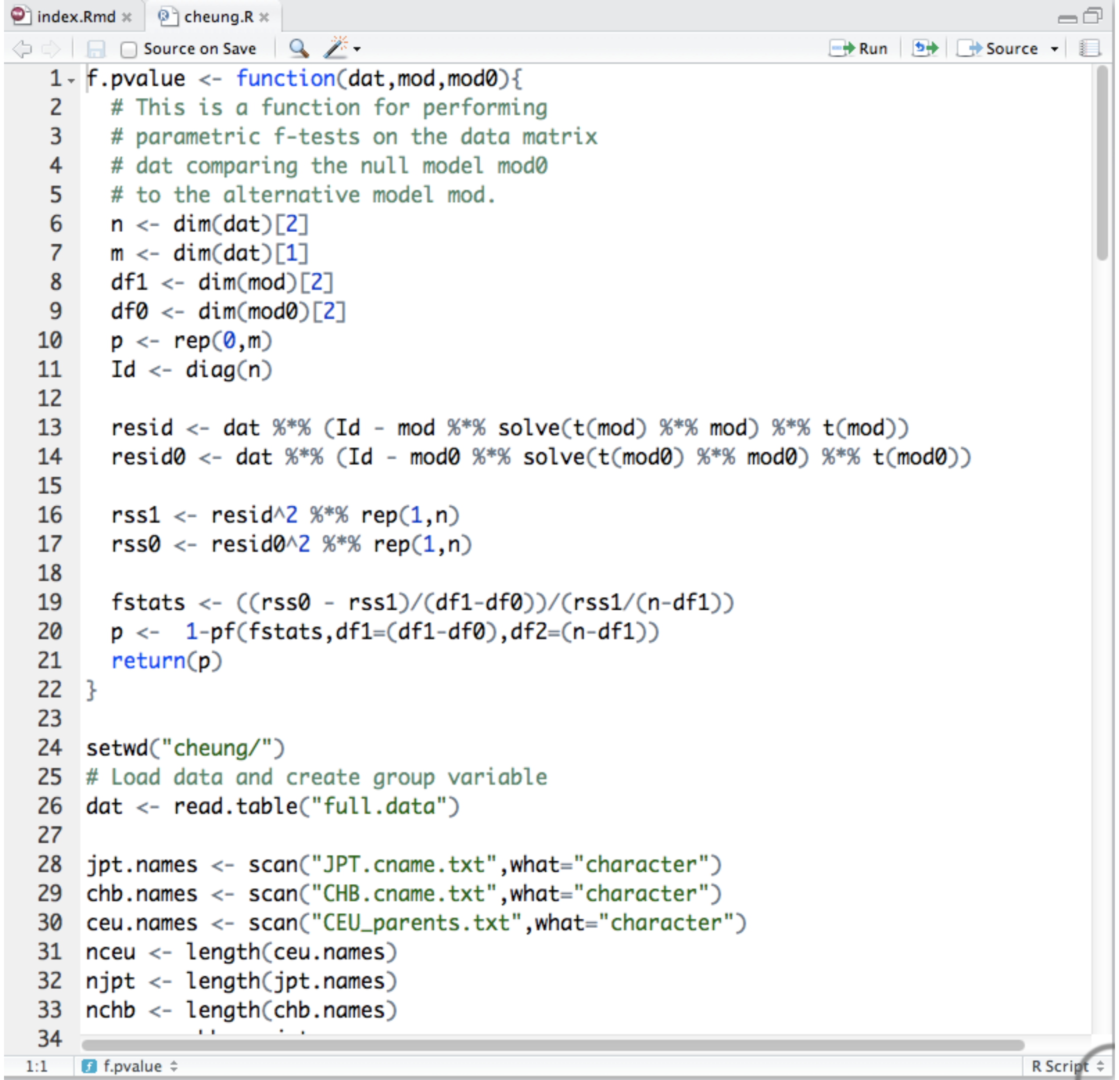
# How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

Reference: https://github.com/jtleek/datasharing

- Code
    - Raw
    - Literate programming



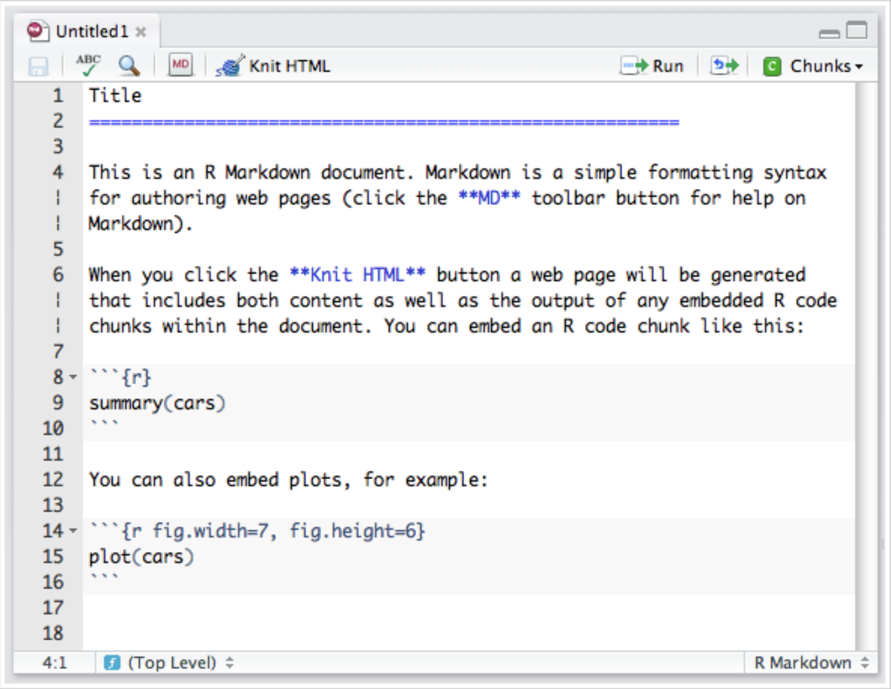
```r
f.pvalue <- function(dat,mod,mod0){
  # This is a function for performing
  # parametric f-tests on the data matrix
  # dat comparing the null model mod0
  # to the alternative model mod.
  n <- dim(dat)[2]
  m <- dim(dat)[1]
  df1 <- dim(mod)[2]
  df0 <- dim(mod0)[2]
  p <- rep(0,m)
  Id <- diag(n)

  resid <- dat %*% (Id - mod %*% solve(t(mod) %*% mod) %*% t(mod))
  resid0 <- dat %*% (Id - mod0 %*% solve(t(mod0) %*% mod0) %*% t(mod0))

  rss1 <- resid^2 %*% rep(1,n)
  rss0 <- resid0^2 %*% rep(1,n)

  fstats <- ((rss0 - rss1)/(df1-df0))/(rss1/(n-df1))
  p <-  1-pf(fstats,df1=(df1-df0),df2=(n-df1))
  return(p)
}

setwd("cheung/")
# Load data and create group variable
dat <- read.table("full.data")

jpt.names <- scan("JPT.cname.txt",what="character")
chb.names <- scan("CHB.cname.txt",what="character")
ceu.names <- scan("CEU_parents.txt",what="character")
nceu <- length(ceu.names)
njpt <- length(jpt.names)
nchb <- length(chb.names)
```

R Markdown Documents

http://rmarkdown.rstudio.com/

To work with R Markdown (.Rmd) files in RStudio you first need to ensure that the knitr package (version 0.5 or later) is installed.

To create a new R Markdown file, go to **File | New |** and select **R Markdown**. A new file is create with a default template to get you oriented:



~~~
Title
========================================================

This is an R Markdown document. Markdown is a simple formatting syntax
for authoring web pages (click the **MD** toolbar button for help on
Markdown).

When you click the **Knit HTML** button a web page will be generated
that includes both content as well as the output of any embedded R code
chunks within the document. You can embed an R code chunk like this:

```{r}
summary(cars)
~~~

You can also embed plots, for example:

```{r fig.width=7, fig.height=6}
plot(cars)
```

```

Note that the toolbar provides some useful tools for working with R Markdown:

- **Quick Reference** — Click the **MD** toolbar button to open a quick reference guide for Markdown.
- **Knit HTML** — Click to knit the current document to HTML, see the **Knitting to HTML** section below for more details.
- **Run** — Run the current line or selection of lines in the console. This allows running R code inside a code chunk similar to a normal R source file.
- **Chunks** — The chunks menu provides assistance with inserting, running, and chunk navigation. See the **Chunk Menu and Options** section below for more details.
```