KEN BUTLER

PROBLEMS AND SOLUTIONS IN APPLIED STATISTICS

Contents

1
Introduction

This book will hold a collection of problems, and my solutions to them, in applied statistics with R. These come from my courses STAC32 and STAD29 at the University of Toronto Scarborough.

The problems were originally written in Sweave (that is, LaTeX with R code chunks), using the `exam` document class, using data sets stolen from numerous places (textbooks, websites etc). I wrote a Perl program to strip out the LaTeX and turn each problem into R Markdown for this book. You will undoubtedly see bits of LaTeX still embedded in the text. I am trying to update my program to catch them, but I am sure to miss some. If you see anything, file an issue on the Github page for now. I want to fix problems programmatically at first, but when the majority of the problems have been caught, I will certainly take pull requests. I will acknowledge all the people who catch things.

working on stuff from assignments 9/9a

look at heat data

rejig the crickets questions so less duplication

bodyfat and bodyfat-sign duplication

2
Getting used to R and R Studio

We begin with this:

```
library(tidyverse)
```

```
## -- Attaching packages ---- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

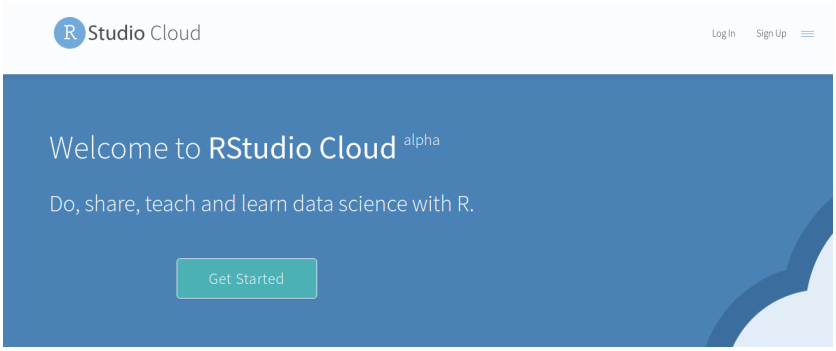
and so to the problems:

2.1 *Getting an R Studio Cloud account*

Follow these steps to get an R Studio Cloud account.

(a) Point your web browser at link. (If you already have R and R Studio installed on your computer, you can use that instead, throughout the course; just do part (d) of this question. Any references to R Studio Cloud in this assignment also apply to R Studio on your computer.)

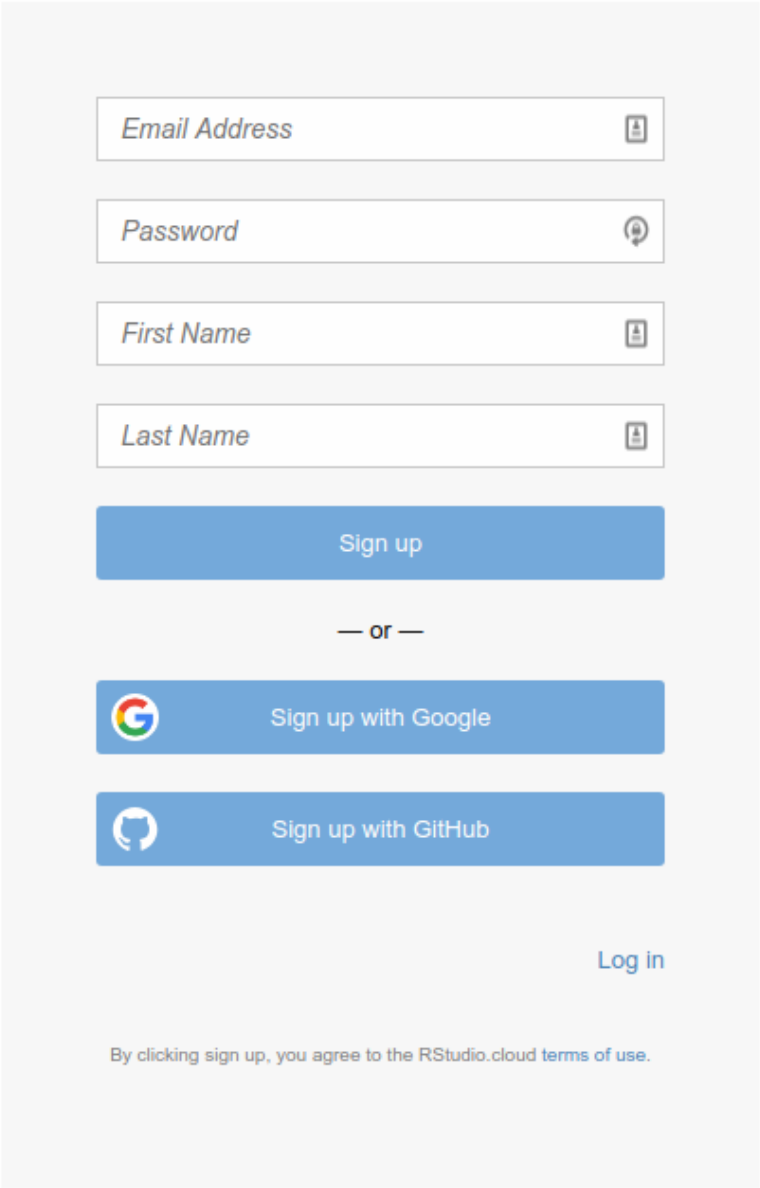
Solution
You should see this:



. Click on Get Started. You might instead see the screen in the next part.

(b) Choose an account to use.

Solution
Here's what you should see now:

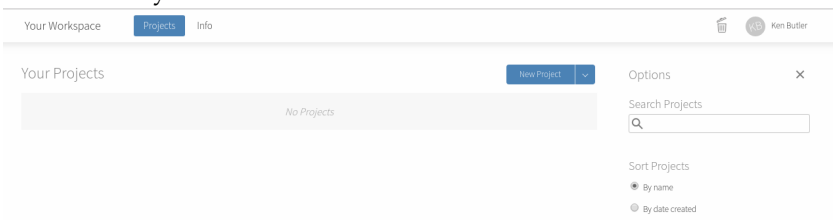


If you're happy with using your Google account, click that button. You will probably have to enter your Google password. (If you are doing this on your own computer, you might not have to do that.) If you have a GitHub account and you want to use *that*, same principle. You can also use an email address as your login to R Studio Cloud. (You can use any e-mail address; I'm not checking.) Enter it in the top box, and enter a password to use with R Studio Cloud in the second. (This does not have to be, and indeed probably should not be, the same as your email password.) Below that, enter your first and last name. This will appear at the top right of the screen when

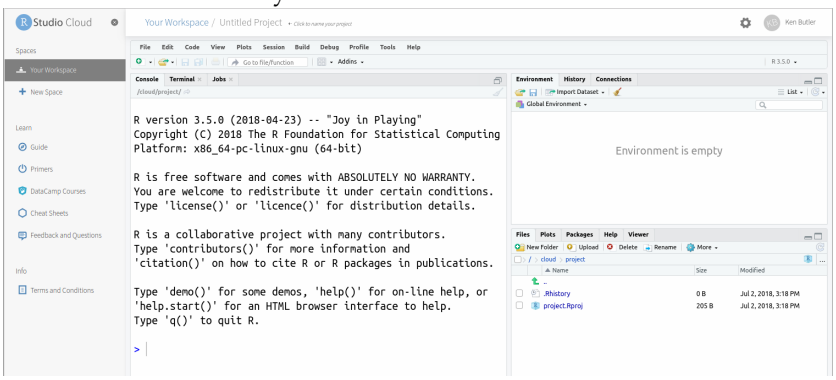
you are logged in. Then click Sign Up. After that, you will have to make a unique account name (which *you* actually never use, but which `rstudio.cloud` uses to name your files). After that, you are automatically logged in.

(c) Take a look around, and create a new Project. Give the new project any name you like.

Solution
This is what you see now:



Click on the blue New Project button to create a new Project. (A project is a self-contained piece of work, like for example an assignment.) You will see the words ''Loading Project'' and spinning circles for a few moments. Then you see this:



To give your project a name, click at the top where it says Untitled Project and type a name like Assignment 0 into the box.

(d) Before we get to work, look for the blue > at the bottom left. Click next to it to get a flashing cursor, and then type what you see here (in blue):

```
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.

> install.packages("tidyverse")
```

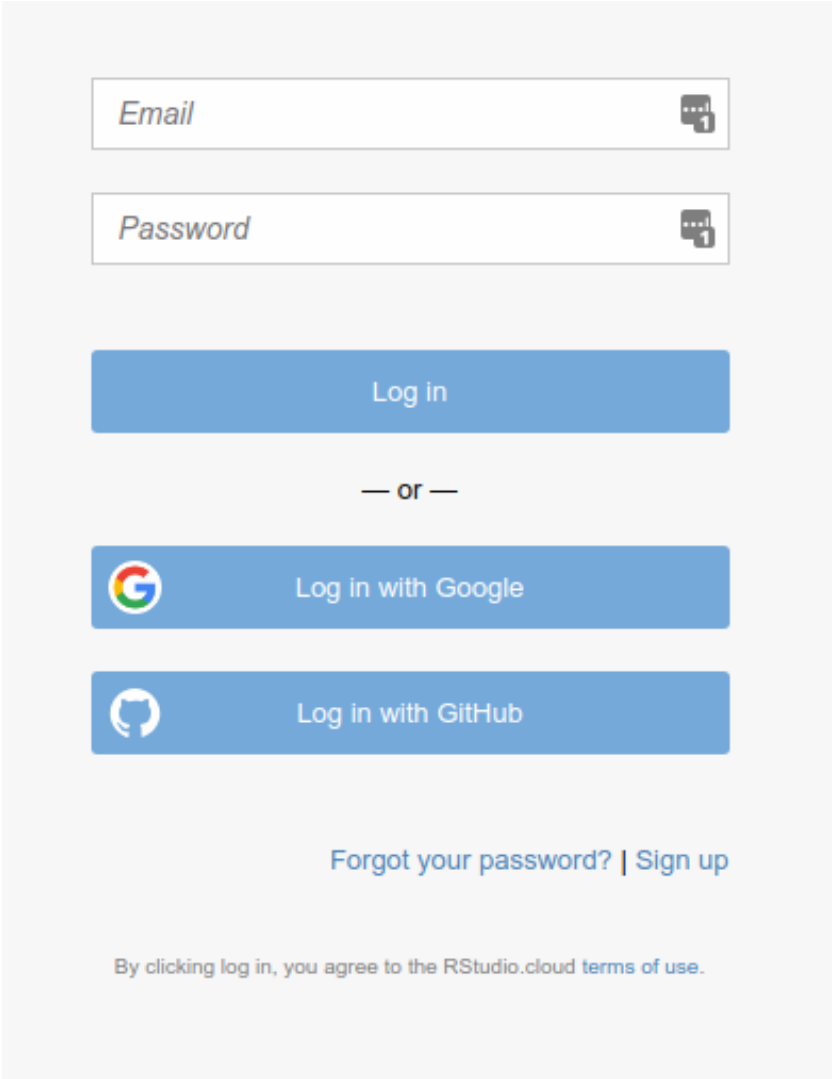
Then press Enter.
Solution

This lets it install a bunch of things. It may take some time. If you are watching it, look out for lines beginning with `g++`, which are C++ code that needs to be compiled. This is the end of what I had. Look out for the word DONE near the bottom:

```
** building package indices
** installing vignettes
** testing if installed package can be loaded
* DONE (dplyr)
* installing *binary* package ‘dbplyr’ ...
* DONE (dbplyr)
* installing *binary* package ‘tidyr’ ...
* DONE (tidyr)
* installing *binary* package ‘broom’ ...
* DONE (broom)
* installing *binary* package ‘modelr’ ...
* DONE (modelr)
* installing *binary* package ‘tidyverse’ ...
* DONE (tidyverse)

The downloaded source packages are in
        ‘/tmp/Rtmp8xSm43/downloaded_packages’
>
```

(e) Not for now, but for later: if you are on a lab computer, you should probably log out when you are done. To do that, find your name at the top right. Click on it, and two things should pop out to the right: Profile and Log Out. Select Log Out. You should be returned to one of the screens you began with, possibly the Welcome to R Studio Cloud one. To log back in, now or next time, look for Log In at the top right. Click it, to get this:



and then you can log in with your email and password, or Google or Github IDs, whichever you used. Now we can get down to some actual work.

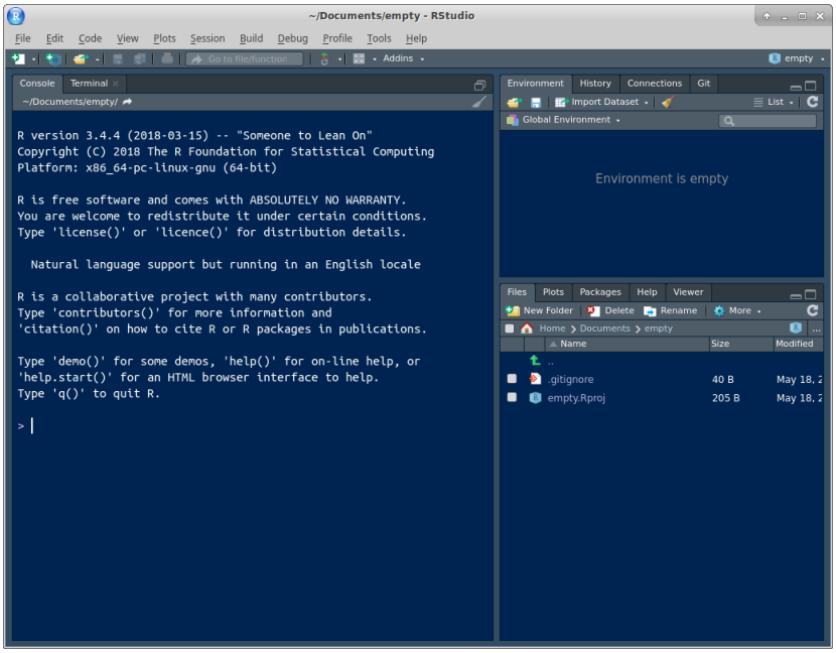
2.2 *Getting started*

This question is to get you started using R.

(a) Start R Studio Cloud, in some project. (If you started up a new project in the previous question and are still logged in, use that; if not, create a new project.)

Solution

You ought to see something like this. I have a dark blue background here, which you probably do not.

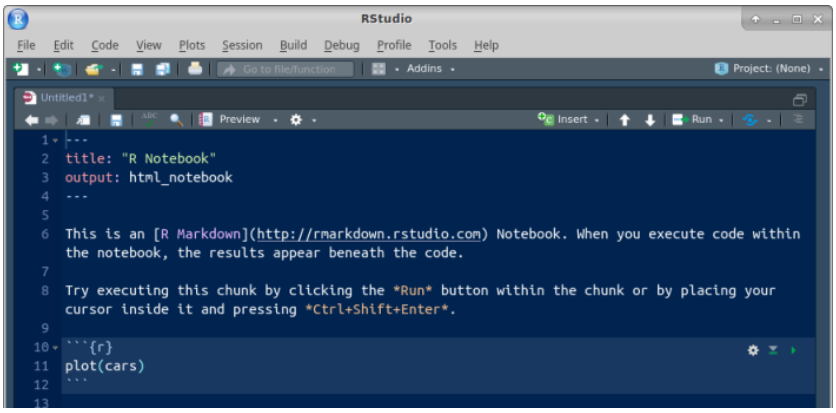


It won't look exactly like that (for example, the background will probably be white) but there should be one thing on the left half, and at the top right it'll say ''Environment is empty''. Extra: if you want to tweak things, select Tools (at the top of the screen) and from it Global Options, then click Appearance. You can make the text bigger or smaller via Editor Font Size, and choose a different colour scheme by picking one of the Editor Themes (which previews on the right). My favourite is Tomorrow Night Blue. Click Apply or OK when you have found something you like. (I spend a lot of time in R Studio, and I like having a dark background to be easier on my eyes.)

(b) We're going to do some stuff in R here, just to get used to it. First, make an R Notebook by selecting File, New File and R Notebook.

Solution

The first time, you'll be invited to ''install some packages'' to make the Notebook thing work. Let it do that by clicking Yes. After that, you'll have this:

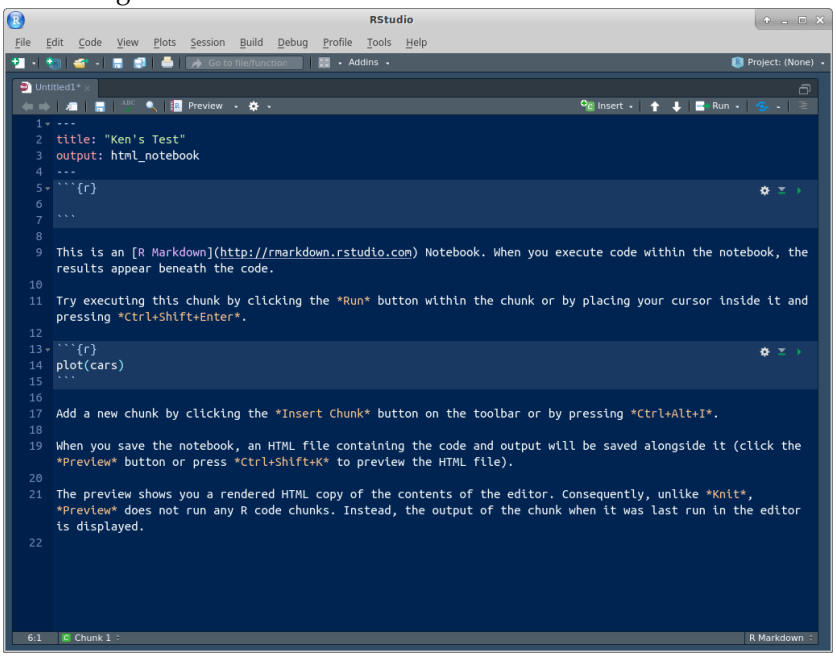


Find the Insert and Run buttons along the top of the R Notebook window. We'll be using them shortly. (The template notebook may or may not be maximized; it doesn't matter either way. You might see all four panes or as few as one. If you want to control that, select View at the top, then Panes, then either Show All Panes or Zoom Source, as you prefer. In the menus, you'll also see keyboard shortcuts for these, which you might find worth learning.)

(c) Change the title to something of your choosing. Then go down to line 5, click on the Insert button and select R. You should see a ''code chunk'' appear at line 5, which we are going to use in a moment.

Solution

Something like this:



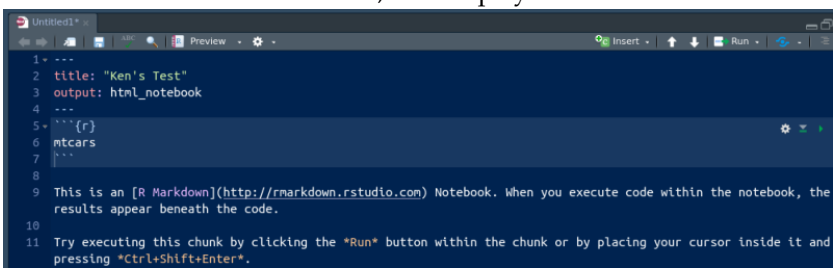
(d) Type the line of code shown below into the chunk in the R Note-

book:

```
mtcars
```

Solution

What this will do: get hold of a built-in data set with information about some different models of car, and display it.

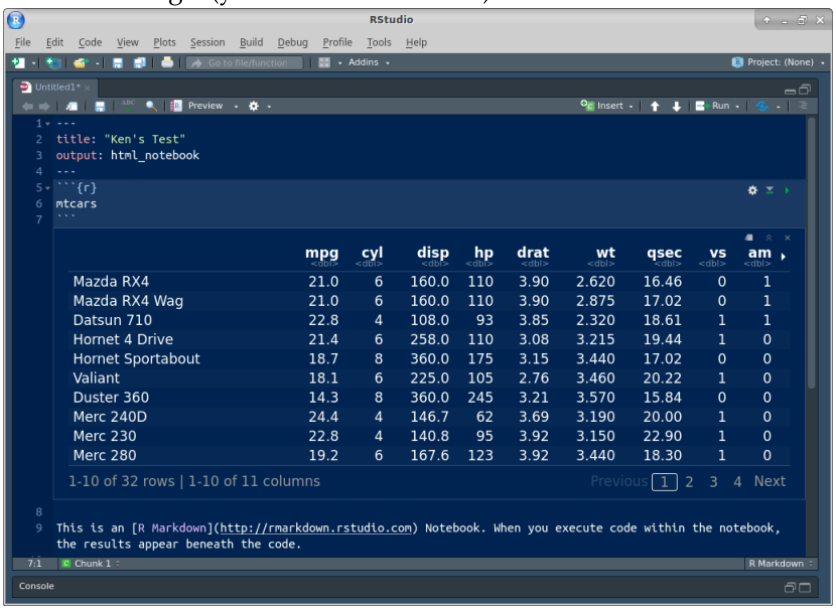


In approximately five seconds, you'll be demonstrating that for yourself.

(e) Run this command. To do that, look at the top right of your code chunk block (shaded in a slightly different colour). You should see a gear symbol, a down arrow and a green ''play button''. Click the play button. This will run the code, and show the output below the code chunk.

Solution

Here's what I get (yours will be the same).



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0

1-10 of 32 rows | 1-10 of 11 columns Previous 1 2 3 4 Next

This is a rectangular array of rows and columns, with individuals in rows and variables in columns, known as a ''data frame''. When you display a data frame in an R Notebook, you see 10 rows and as many

columns as will fit on the screen. At the bottom, it says how many rows and columns there are altogether (here 32 rows and 11 columns), and which ones are being displayed. You can see more rows by clicking on Next, and if there are more columns, you'll see a little arrow next to the rightmost column (as here next to `am`) that you can click on to see more columns. Try it and see. Or if you want to go to a particular collection of rows, click one of the numbers between Previous and Next: 1 is rows 1–10, 2 is rows 11–20, and so on. The column on the left without a header (containing the names of the cars) is called ''row names''. These have a funny kind of status, kind of a column and kind of not a column; usually, if we need to use the names, we have to put them in a column first. In future solutions, rather than showing you a screenshot, expect me to show you something like this:

```
mtcars
```

```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec
##  * <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  21       6  160    110  3.9   2.62  16.5
##  2  21       6  160    110  3.9   2.88  17.0
##  3  22.8     4  108     93  3.85  2.32  18.6
##  4  21.4     6  258    110  3.08  3.22  19.4
##  5  18.7     8  360    175  3.15  3.44  17.0
##  6  18.1     6  225    105  2.76  3.46  20.2
##  7  14.3     8  360    245  3.21  3.57  15.8
##  8  24.4     4  147.    62  3.69  3.19  20
##  9  22.8     4  141.    95  3.92  3.15  22.9
## 10  19.2     6  168.   123  3.92  3.44  18.3
## # ... with 22 more rows, and 4 more
## #   variables: vs <dbl>, am <dbl>,
## #   gear <dbl>, carb <dbl>
```

The top bit is the code, the bottom bit with the `##` the output. In this kind of display, you only see the first ten rows (by default).

If you don't see the ''play button'', make sure that what you have really is a code chunk. (I often accidentally delete one of the special characters above or below the code chunk). If you can't figure it out, delete this code chunk and make a new one. Sometimes R Studio gets confused.

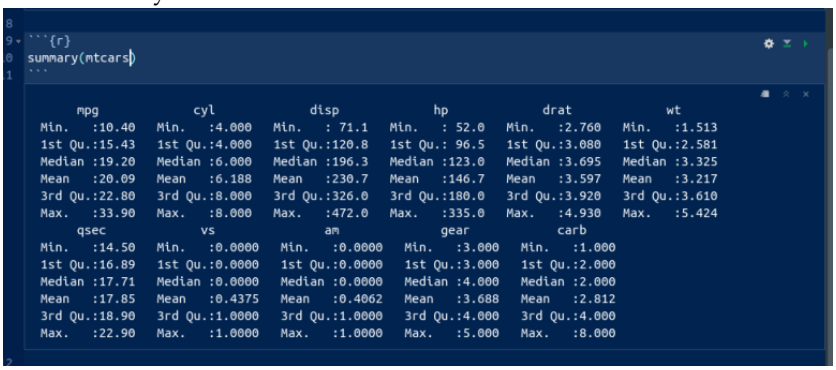
On the code chunk, the other symbols are the settings for this chunk (you have the choice to display or not display the code or the output or to not actually run the code). The second one, the down arrow, runs all the chunks prior to this one (but not this one).

The output has its own little buttons. The first one pops the output out into its own window; the second one shows or hides the out-

put, and the third one deletes the output (so that you have to run the chunk again to get it back). Experiment. You can't do much damage here.

(f) Something a little more interesting: `summary` obtains a summary of whatever you feed it (the five-number summary plus the mean for numerical variables). Obtain this for our data frame. To do this, create a new code chunk below the previous one, type `summary(mtcars)` into the code chunk, and run it.

Solution
This is what you should see:



or the other way:

```r
summary(mtcars)
```

```
##       mpg             cyl
##  Min.   :10.40   Min.   :4.000
##  1st Qu.:15.43   1st Qu.:4.000
##  Median :19.20   Median :6.000
##  Mean   :20.09   Mean   :6.188
##  3rd Qu.:22.80   3rd Qu.:8.000
##  Max.   :33.90   Max.   :8.000
##       disp             hp
##  Min.   : 71.1   Min.   : 52.0
##  1st Qu.:120.8   1st Qu.: 96.5
##  Median :196.3   Median :123.0
##  Mean   :230.7   Mean   :146.7
##  3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :472.0   Max.   :335.0
##       drat             wt
##  Min.   :2.760   Min.   :1.513
##  1st Qu.:3.080   1st Qu.:2.581
##  Median :3.695   Median :3.325
##  Mean   :3.597   Mean   :3.217
##  3rd Qu.:3.920   3rd Qu.:3.610
```