

Hacking Religion: TRS & Data Science in Action

Jeremy H. Kidwell

2023-09-29

Table of contents

Preface	3
1 Introduction: Hacking Religion	4
1.1 Who this book is for	4
1.2 Why this book?	4
1.3 The hacker way	4
1.4 Why programmatic data science?	4
1.5 Learning to code: my way	5
1.6 Getting set up	6
2 The 2021 UK Census	8
2.1 Your first project: building a pie chart	8
2.1.1 Examining data:	9
2.1.2 Parsing and Exploring your data	11
2.2 Making your first chart	11
References	15
3 Survey Data: Spotlight Project	16
References	18
4 Mapping churches: geospatial data science	19
References	20
5 Data scraping, corpus analysis and wordclouds	21
References	22
6 Summary	23
References	24

Preface

This is a Quarto book.

To learn more about Quarto books visit <https://quarto.org/docs/books>.

1 Introduction: Hacking Religion

1.1 Who this book is for

1.2 Why this book?

1.3 The hacker way

1. Tell the truth
2. Do not deceive using beauty
3. Work transparently: research as open code using open data
4. Draw others in: produce reproducible research
5. Learn by doing

1.4 Why programmatic data science?

This isn't just a book about data analysis, I'm proposing an approach which might be thought of as research-as-code, where you write out instructions to execute the various steps of work. The upside of this is that other researchers can learn from your work, correct and build on it as part of the commons. It takes a bit more time to learn and set things up, but the upside is that you'll gain access to a set of tools and a research philosophy which is much more powerful.

1.5 Learning to code: my way

This guide is a little different from other textbooks targetting learning to code. I remember when I was first starting out, I went through a fair few guides, and they all tended to spend about 200 pages on various theoretical bits, how you form an integer, or data structures, subroutines, or whatever, such that it was weeks before I got to actually *do* anything. I know some people, may prefer this approach, but I dramatically prefer a problem-focussed approach to learning. Give me something that is broken, or a problem to solve, which engages the things I want to figure out and the motivation for learning just comes much more naturally. And we know from research in cognitive science that these kinds of problem-focussed approaches can tend to facilitate faster learning and better retention, so it's not just my personal preference, but also justified! It will be helpful for you to be aware of this approach when you get into the book as it explains some of the editorial choices I've made and the way I've structured things. Each chapter focusses on a *problem* which is particularly salient for the use of data science to conduct research into religion. That problem will be my focal point, guiding choices of specific aspects of programming to introduce to you as we work our way around that data set and some of the crucial questions that arise in terms of how we handle it. If you find this approach unsatisfying, luckily there are a number of really terrific guides which lay things out slowly and methodically and I will explicitly signpost some of these along the way so that you can do a "deep dive" when you feel like it. Otherwise, I'll take an accelerated approach to this introduction to data science in R. I expect that you will identify adjacent resources and perhaps even come up with your own creative approaches along the way, which incidentally is how real data science tends to work in practice.

There are a range of terrific textbooks out there which cover all these elements in greater depth and more slowly. In particular, I'd recommend that many readers will want to check out Hadley Wickham's "R For Data Science" book. I'll include marginal notes in this guide pointing to sections of that book, and a few others which unpack the basic mechanics of R in more detail.

1.6 Getting set up

Every single tool, programming language and data set we refer to in this book is free and open source. These tools have been produced by professionals and volunteers who are passionate about data science and research and want to share it with the world, and in order to do this (and following the “hacker way”) they’ve made these tools freely available. This also means that you aren’t restricted to a specific proprietary, expensive, or unavailable piece of software to do this work. I’ll make a few opinionated recommendations here based on my own preferences and experience, but it’s really up to your own style and approach. In fact, given that this is an open source textbook, you can even propose additions to this chapter explaining other tools you’ve found that you want to share with others.

There are, right now, primarily two languages that statisticians and data scientists use for this kind of programmatic data science: python and R. Each language has its merits and I won’t rehash the debates between various factions. For this book, we’ll be using the R language. This is, in part, because the R user community and libraries tend to scale a bit better for the work that I’m commending in this book. However, it’s entirely possible that one could use python for all these exercises, and perhaps in the future we’ll have volume two of this book outlining python approaches to the same operations.

Bearing this in mind, the first step you’ll need to take is to download and install R. You can find instructions and install packages for a wide range of hardware on the The Comprehensive R Archive Network (or “CRAN”): <https://cran.rstudio.com>. Once you’ve installed R, you’ve got some choices to make about the kind of programming environment you’d like to use. You can just use a plain text editor like `textedit` to write your code and then execute your programs using the R software you’ve just installed. However, most users, myself included, tend to use an integrated development environment (or “IDE”). This is usually another software package with a guided user interface and some visual elements that make it faster to write and test your code. Some IDE packages, will have built-in reference tools so you can

look up options for libraries you use in your code, they will allow you to visualise the results of your code execution, and perhaps most important of all, will enable you to execute your programs line by line so you can spot errors more quickly (we call this “debugging”). The two most popular IDE platforms for R coding at the time of writing this textbook are RStudio and Visual Studio. You should download and try out both and stick with your favourite, as the differences are largely aesthetic. I use a combination of RStudio and an enhanced plain text editor Sublime Text for my coding.

Once you have R and your pick of an IDE, you are ready to go! Proceed to the next chapter and we’ll dive right in and get started!

2 The 2021 UK Census

2.1 Your first project: building a pie chart

Let's start by importing some data into R. Because R is what is called an object-oriented programming language, we'll always take our information and give it a home inside a named object. There are many different kinds of objects, which you can specify, but usually R will assign a type that seems to fit best.

In the example below, we're going to read in data from a comma separated value file ("csv") which has rows of information on separate lines in a text file with each column separated by a comma. This is one of the standard plain text file formats. R has a function you can use to import this efficiently called "read.csv". Each line of code in R usually starts with the object, and then follows with instructions on what we're going to put inside it, where that comes from, and how to format it:

If you'd like to explore this all in a bit more depth, you can find a very helpful summary in R for Data Science, chapter 8, "[data import](#)".

```
# R Setup -----  
setwd("/Users/kidwellj/gits/hacking_religion_textbook/hacking_religion")  
library(here) # much better way to manage working paths in R across multiple instances
```

here() starts at /Users/kidwellj/gits/hacking_religion_textbook

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --  
v dplyr      1.1.3      v readr      2.1.4  
v forcats    1.0.0      v stringr    1.5.0  
v ggplot2    3.4.3      v tibble     3.2.1  
v lubridate  1.9.3      v tidyr      1.3.0  
v purrr      1.0.2
```



```
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
here::i_am("chapter_1.qmd")
```

here() starts at /Users/kidwellj/gits/hacking_religion_textbook/hacking_religion

```
religion_uk <- read.csv(here("example_data", "census2021-ts030-rgn.csv"))
```

2.1.1 Examining data:

What's in the table? You can take a quick look at either the top of the data frame, or the bottom using one of the following commands:

```
head(religion_uk)
```

	geography	total	no_religion	christian	buddhist	hindu	jewish
1	North East	2647012	1058122	1343948	7026	10924	4389
2	North West	7417397	2419624	3895779	23028	49749	33285
3	Yorkshire and The Humber	5480774	2161185	2461519	15803	29243	9355
4	East Midlands	4880054	1950354	2214151	14521	120345	4313
5	West Midlands	5950756	1955003	2770559	18804	88116	4394
6	East of England	6335072	2544509	2955071	26814	86631	42012
	muslim	sikh	other	no_response			
1	72102	7206	9950	133345			
2	563105	11862	28103	392862			
3	442533	24034	23618	313484			
4	210766	53950	24813	286841			
5	569963	172398	31805	339714			
6	234744	24284	36380	384627			

This is actually a fairly ugly table, so I'll use an R tool called kable to give you prettier tables in the future, like this:

```
knitr::kable(head(religion_uk))
```

geography	total	no_religion	christian	hindu	jewish	muslim	sikh	other	no_response
North East	2647010	105812	234394	7826	10924	43897	21072	20699	50133345
North West	7417327	19623	489572	30284	9743	32856	31058	62810	392862
Yorkshire and The Humber	5480774	11824	46151	9803	29243	35544	2521	3034	2361813484
East Midlands	4880059	50352	21415	14521	11203	4513	21076	3950	2481386841
West Midlands	5950756	50027	70518	8048	8116	4394	56996	37239	3180339714
East Midlands	6335072	44502	29550	2681	4866	3420	12347	2428	4638084627

You can see how I've nested the previous command inside the `kable` command. For reference, in some cases when you're working with really complex scripts with many different libraries and functions, they may end up with functions that have the same name. You can specify the library where the function is meant to come from by preceding it with `::` as we've done `knitr::` above. The same kind of output can be gotten using `tail`:

```
knitr::kable(tail(religion_uk))
```

geography	total	no_religion	christian	hindu	jewish	muslim	sikh	other	no_response
5 West Midlands	5950756	50027	70518	8048	8116	4394	56996	37239	3180339714
6 East Midlands	6335072	44502	29550	2681	4866	3420	12347	2428	4638084627
7 London	8799728	80403	57768	7425	4530	3454	6618	7445	4675015662
8 South East	9278058	33094	31335	4433	15474	1868	20906	7434	5409566279

	geography	total	no_religion	christian	hindu	islamic	jewish	muslim	sikh	other	no_response
9	South West	5701186	1351	3362	2635	821	579	2774	673	8780	15274
10	Wales	3107494	146	398	835	477	130	751	224	2044	66947

2.1.2 Parsing and Exploring your data

The first thing you’re going to want to do is to take a smaller subset of a large data set, either by filtering out certain columns or rows. Now let’s say we want to just work with the data from the West Midlands, and we’d like to omit some of the columns. We can choose a specific range of columns using `select`, like this:

You can use the `filter` command to do this. To give an example, `filter` can pick a single row in the following way:

```
wmids_data <- religion_uk %>%
  filter(geography=="West Midlands")
```

Now we’ll use `select` in a different way to narrow our data to specific columns that are needed (no totals!).

In keeping with my goal to demonstrate data science through examples, we’re going to move on to producing some snappy looking charts for this data.

Some readers will want to pause here and check out Hadley Wickham’s “R For Data Science” book, in the section, “[Data visualisation](#)” to get a fuller explanation of how to explore your data.

2.2 Making your first chart

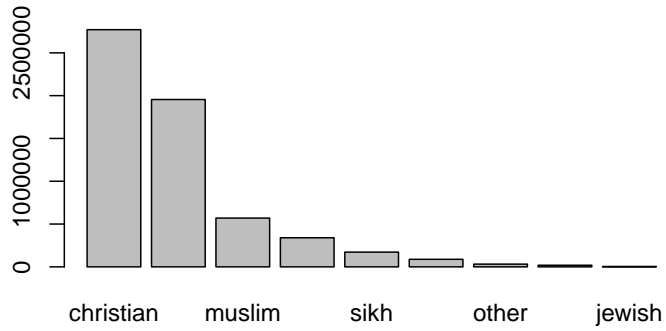
We’ve got a nice lean set of data, so now it’s time to visualise this. We’ll start by making a pie chart:

```
wmids_data <- wmids_data %>% select(no_religion:no_response)
wmids_data <- gather(wmids_data)
```

There are two basic ways to do visualisations in R. You can work with basic functions in R, often called “base R” or you can work with an alternative library called `ggplot`:

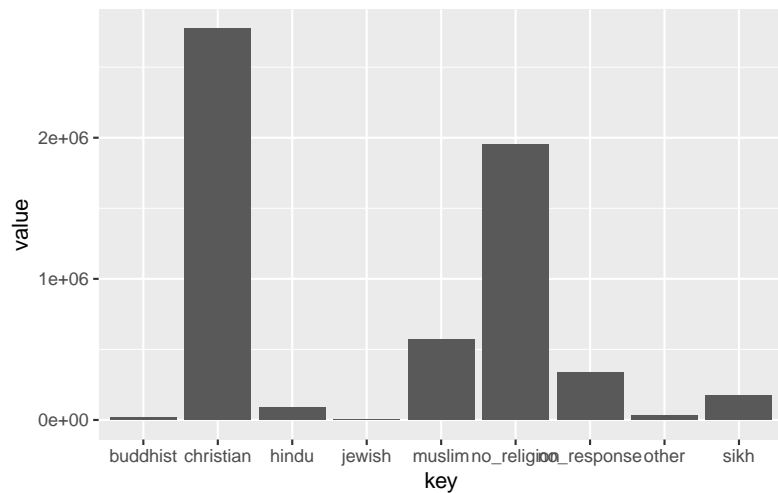
2.2.0.1 Base R

```
df <- wmid_data[order(wmid_data$value,decreasing = TRUE),]  
barplot(height=df$value, names=df$key)
```

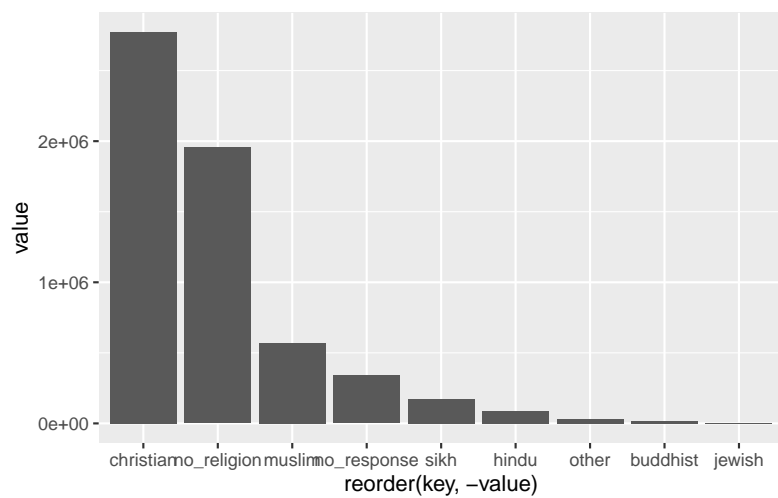


2.2.0.2 GGPlot

```
# unsorted  
ggplot(wmid_data, aes(x = key, y = value)) +  
  geom_bar(stat = "identity")
```



```
# with sorting added in
ggplot(wmids_data, aes(x= reorder(key, -value), value)) + geom_bar(stat = "identity")
```



Clean up chart features

Add time series data for 2001 and 2011 census, change to grouped bar plot:

<https://r-graphics.org/recipe-bar-graph-grouped-bar#discussion-8>

💡 What is Religion?

Content tbd

💡 Hybrid Religious Identity

Content tbd

💡 What is Secularisation?

Content tbd

References

3 Survey Data: Spotlight Project


In the last chapter we explored some high level data about religion in the UK. This was a census sample, which usually refers to an attempt to get as comprehensive a sample as possible. But this is actually fairly unusual in practice. Depending on how complex a subject is, and how representative we want our data to be, it's much more common to use selective sampling, that is survey responses at $n=100$ or $n=1000$ at a maximum. The advantage of a census sample is that you can explore how a wide range of other factors - particularly demographics - intersect with your question. And this can be really valuable in the study of religion, particularly as you will see as we go along that responses to some questions are more strongly correlated to things like economic status or educational attainment than they are to religious affiliation. It can be hard to tell if this is the case unless you have enough of a sample to break down into a number of different kinds of subsets. But census samples are complex and expensive to gather, so they're quite rare in practice.

For this chapter, I'm going to walk you through a data set that a colleague (Charles Ogunbode) and I collected in 2021. Another problem with smaller, more selective samples is that researchers can often undersample minoritised ethnic groups. This is particularly the case with climate change research. Until the time we conducted this research, there had not been a single study investigating the specific experiences of people of colour in relation to climate change in the UK. Past researchers had been content to work with large samples, and assumed that if they had done 1000 surveys and 50 of these were completed by people of colour, they could "tick" the box. But 5% is actually well below levels of representation in the UK generally, and even more sharply the case for specific communities. And

if we bear in mind that non-white respondents are (of course!) a highly heterogenous group, we're even more behind in terms of collecting data that can improve our knowledge. Up until recently researchers just haven't been paying close enough attention to catch the significant neglect of the empirical field that this represents.

While I've framed my comments above in terms of climate change research, it is also the case that, especially in diverse societies like the USA, Canada, the UK etc., paying attention to non-majority groups and people and communities of colour automatically draws in a strongly religious sample. This is highlighted in one recent study done in the UK, the "[Black British Voices Report](#)" in which the researchers observed that "84% of respondents described themselves as religious and/or spiritual". My comments above in terms of controlling for other factors remains important here - these same researchers also note that "despire their significant important to the lives of Black Britons, only 7% of survey respondents reported that their religion was more defining of their identity than their race".

We've decided to open up access to our data and I'm highlighting it in this book because it's a unique opportunity to explore a dataset that emphasises diversity from the start, and by extension, provides some really interesting ways to use data science techniques to explore religion in the UK.

 How can we measure religion?

Content tbd

References

4 Mapping churches: geospatial data science

Guides to geographies: <https://rconsortium.github.io/censusguide/>
<https://ocsi.uk/2019/03/18/lsoas-leps-and-lookups-a-beginners-guide-to-statistical-geographies/>

References

5 Data scraping, corpus analysis and wordclouds

References

6 Summary

An open textbook introducing data science to religious studies

References