Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 02, 2024

Abstract

This report provides an analysis of data related to health, age, socio-economic factors and lifestyle habits in adults (from the age of 16) from the population in England, derived from the Health Survey for England 2019.

Introduction

This is a body of text. *This is an italic body of text.* This is a clickable link!.

Some YAML Stuff

The lion's share of a R Markdown document will be raw text, though the front matter may be the most important part of the document. R Markdown uses YAML for its metadata and the fields differ from what an author would use for a Beamer presentation. I provide a sample YAML metadata largely taken from this exact document and explain it below.

```yaml
---
output:
  pdf_document:
    keep_tex: true
    fig_caption: true
    latex_engine: pdflatex
title: "A Pandoc Markdown Article Starter and Template"
abstract: "This document provides an introduction to R Markdown, argues for its..."
date: "`r format(Sys.time(), '%B %d, %Y')`"
geometry: margin=1in
fontsize: 11pt
# spacing: double
---
```

`output:` will tell R Markdown we want a PDF document rendered with LaTeX. Since we are adding a fair bit of custom options to this call, we specify `pdf_document:` on the next line (with, importantly, a two-space indent). We specify additional output-level options underneath it, each are indented with four spaces. The line (`keep_tex: true`) tells R Markdown to render a raw `.tex` file along with the PDF document. This is useful for both debugging and the publication stage. The next line `fig_caption: true` tells R Markdown to make sure that whatever images are included in the document are treated as figures in which our caption in brackets in a Markdown call is treated as the caption in the figure. The next line (`latex_engine: pdflatex`) tells R Markdown to use pdflatex and not some other option like `lualatex`. For this template, I'm pretty sure this is mandatory.[^pdflatex]

The next fields get to the heart of the document itself. `title:` is, intuitively, the title of the manuscript. Do note that fields like `title:` do not have to be in quotation marks, but must be in quotation marks if the title of the document includes a colon. That said, the only reason to use a colon in an article title is if it is followed by a subtitle, hence the optional field (`subtitle:`). Notice I "comment out" the subtitle in the above example with a pound sign since this particular document does not have a subtitle.

`date` comes standard with R Markdown and you can use it to enter the date of the most recent compile.

The next items are optional and cosmetic. `geometry:` is a standard option in LaTeX. I set the margins at one inch, and you probably should too. `fontsize:` sets, intuitively, the font size. The default is 10-point, but I prefer 11-point. `spacing:` is an optional field. If it is set as "double", the ensuing document is double-spaced. "single" is the only other valid entry for this field, though not including the entry in the YAML metadata amounts to singlespacing the document by default. Notice I have this “commented out” in the example code.

Getting Started with Markdown Syntax

There are a lot of cheatsheets and reference guides for Markdown (e.g. Adam Prichard, Assemble, Rstudio, Rstudio again, Scott Boms, Daring Fireball, among, I’m sure, several others).

```
# Introduction

**Lorem ipsum** dolor *sit amet*.

- Single asterisks italicize text *like this*.
- Double asterisks embolden text **like this**.

Start a new paragraph with a blank line separating paragraphs.

- This will start an unordered list environment, and this will be the first item.
- This will be a second item.
- A third item.
    - Four spaces and a dash create a sublist and this item in it.
- The fourth item.

1. This starts a numerical list.
2. This is no. 2 in the numerical list.

# This Starts A New Section
## This is a Subsection
### This is a Subsubsection
#### This starts a Paragraph Block.

> This will create a block quote, if you want one.

Want a table? This will create one.

Table Header  | Second Header
------------- | -------------
Table Cell    | Cell 2
Cell 3        | Cell 4

Note that the separators *do not* have to be aligned.

Want an image? This will do it.

![caption for my image](path/to/image.jpg)

`fig_caption: yes` will provide a caption. Put that in the YAML metadata.
```

```
Almost forgot about creating a footnote.[^1] This will do it again.[^2]

[^1]: The first footnote
[^2]: The second footnote

Want to cite something?

- Find your biblatexkey in your bib file.
- Put an @ before it, like @smith1984, or whatever it is.
- @smith1984 creates an in-text citation (e.g. Smith (1984) says...)
- [@smith1984] creates a parenthetical citation (Smith, 1984)

That'll also automatically create a reference list at the end of the document.

[In-text link to Google](http://google.com) as well.
```

Exploring the Data

Checking for Messy Data

```r
library(haven) # Required to present the summary of labelled data.
load("~/MA30091/Coursework/MA30091/Datasets/hsesub.Rdata") # The dset is called subdat
summary(subdat)
```

```
##     SerialA             Sex           ag16g10          Age35g     
##  Min.   :2900001   Min.   :1.000   Min.   :1.000   Min.   : 1.00  
##  1st Qu.:2903094   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 8.00  
##  Median :2906238   Median :2.000   Median :4.000   Median :12.00  
##  Mean   :2906229   Mean   :1.539   Mean   :4.128   Mean   :11.71  
##  3rd Qu.:2909378   3rd Qu.:2.000   3rd Qu.:6.000   3rd Qu.:16.00  
##  Max.   :2912465   Max.   :2.000   Max.   :7.000   Max.   :22.00  
##                                    NA's   :2095                   
##      wt_int          topqual2        marstatD         qimd19     
##  Min.   :0.3155   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:0.7941   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000  
##  Median :0.8989   Median :3.000   Median :2.000   Median :3.000  
##  Mean   :1.0000   Mean   :3.664   Mean   :2.658   Mean   :3.044  
##  3rd Qu.:1.0974   3rd Qu.:7.000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :6.4927   Max.   :8.000   Max.   :6.000   Max.   :5.000  
##                   NA's   :2141    NA's   :2096                   
##     urban14b        origin2        cigsta3_19      cigdyal_19    
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.: 0.000  
##  Median :1.000   Median :1.000   Median :3.000   Median : 0.000  
##  Mean   :1.181   Mean   :1.343   Mean   :2.437   Mean   : 1.692  
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.: 0.000  
##  Max.   :2.000   Max.   :5.000   Max.   :3.000   Max.   :60.000  
##                  NA's   :33      NA's   :2151    NA's   :2152
##      BMIVal          NDPNow_19        dnoft_19       drinkYN_19
##  Min.   : 9.723   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:21.915   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :25.904   Median :4.000   Median :4.000   Median :2.000
##  Mean   :26.223   Mean   :3.862   Mean   :4.281   Mean   :1.808
##  3rd Qu.:29.953   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:2.000
##  Max.   :73.494   Max.   :4.000   Max.   :8.000   Max.   :2.000
##  NA's   :2224     NA's   :2148    NA's   :3594    NA's   :2146
##    d7many3_19       omsysval          GOR1
##  Min.   :0.000   Min.   : 75.0   Min.   :1.000
##  1st Qu.:0.000   1st Qu.:110.5   1st Qu.:3.000
##  Median :1.000   Median :121.0   Median :5.000
##  Mean   :1.595   Mean   :122.9   Mean   :5.163
##  3rd Qu.:3.000   3rd Qu.:133.5   3rd Qu.:8.000
##  Max.   :7.000   Max.   :209.5   Max.   :9.000
##  NA's   :2147    NA's   :5593
```

This tells us that all of our variables are coded as numeric. However, we may want to code some as factor variables instead based on the variable descriptions.

- Sex: Should be coded as

Code	Decode	Count
1	Male	
2	Female	
-1	Not Applicable	
-8	Don’t Know	
-9	Refused	

- Age35g: Should be coded as

Code	Decode	Count
1	0-1yrs	
2	2-4yrs	
3	5-7yrs	
4	8-10yrs	
5	11-12yrs	
6	13-15yrs	
7	16-19yrs	
8	20-24yrs	
9	25-29yrs	
10	30-34yrs	
11	35-39yrs	
12	40-44yrs	
13	45-49yrs	
14	50-54yrs	
15	55-59yrs	

Code	Decode	Count
16	60-64yrs	
17	65-69yrs	
18	70-74yrs	
19	75-79yrs	
20	80-84yrs	
21	85-59yrs	
22	90+yrs	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- ag16g10: Should be coded as

Code	Decode	Count
1	16-24yrs	
2	25-34yrs	
3	35-44yrs	
4	45-54yrs	
5	55-64yrs	
6	65-74yrs	
7	75+yrs	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- topqual2: Should be coded as

Code	Decode	Count
1	NVQ4/NVQ5/Degree or equiv	
2	Higher ed below degree	
3	NVQ3/GCE A Level equiv	
4	NVQ2/GCE O Level equiv	
5	NVQ1/CSE other grade equiv	
6	Foreign/other	
7	No qualification	
8	FT Student	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- qimd19: Should be coded as

Code	Decode	Count
1	Most deprived	
5	Least deprived	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

Note: IMD2,IMD3 and IMD4 had no observations.

- urban14b: Should be coded as

Code	Decode	Count
1	Urban	
2	Town/ Fringe/ Village, hamlet and isolated dwellings	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- origin2: Should be coded as

Code	Decode	Count
1	White	
2	Black	
3	Asian	
4	Mixed/multiple ethnic background	
5	Any other ethnic group	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- cigsta3_19: Should be coded as

Code	Decode	Count
1	Current cigarette smoker	
2	Ex-regular cigarette smoker	
3	Never regular cigarette smoker	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- NDPNow_19: Should be coded as

Code	Decode	Count
1	E-cigarettes or vaping devices only	

Code	Decode	Count
2	Other nicotine delivery products only	
3	Both	
4	None	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- drinkYN_19: Should be coded as

Code	Decode	Count
1	No	
2	Yes	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- dnoft_19: Should be coded as

Code	Decode	Count
1	Almost every day	
2	Five or six days a week	
3	Three or four days a week	
4	Once or twice a week	
5	Once or twice a month	
6	Once every couple of months	
7	Once or twice a year	
8	Not at all in the last 12 months	
-1	Not Applicable	
-8	Don't Know	
-9	Refused	

- GOR1: Should be coded as

Code	Decode	Count
1	North East	
2	North West	
3	Yorkshire and the Humber	
4	East Midlands	
5	West Midlands	
6	East of England	
7	London	
8	South East	
9	South West	
-1	Not Applicable	

Code	Decode	Count
-8	Don’t Know	
-9	Refused	

```
subdat$Sex = factor(subdat$Sex)
subdat$Age35g = factor(subdat$Age35g)
subdat$ag16g10 = factor(subdat$ag16g10)
subdat$topqual2 = factor(subdat$topqual2)
subdat$qimd19 = factor(subdat$qimd19)
subdat$urban14b = factor(subdat$urban14b)
subdat$origin2 = factor(subdat$origin2)
subdat$cigsta3_19 = factor(subdat$cigsta3_19)
subdat$NDPNow_19 = factor(subdat$NDPNow_19)
subdat$drinkYN_19 = factor(subdat$drinkYN_19)
subdat$dnoft_19 = factor(subdat$dnoft_19)
subdat$GOR1 = factor(subdat$GOR1)
summary(subdat)
```

```
##     SerialA        Sex         ag16g10         Age35g         wt_int
##  Min.   :2900001   1:4745   4      :1416   14     : 735   Min.   :0.3155
##  1st Qu.:2903094   2:5554   3      :1397   11     : 725   1st Qu.:0.7941
##  Median :2906238            5      :1349   15     : 693   Median :0.8989
##  Mean   :2906229            6      :1242   13     : 681   Mean   :1.0000
##  3rd Qu.:2909378            2      :1083   12     : 672   3rd Qu.:1.0974
##  Max.   :2912465            (Other):1717   16     : 656   Max.   :6.4927
##                             NA's   :2095   (Other):6137
##     topqual2       marstatD     qimd19   urban14b origin2     cigsta3_19
##  1      :2320   Min.   :1.000   1:2074   1:8433   1   :8561   1   :1254
##  7      :1616   1st Qu.:2.000   2:1942   2:1866   2   : 345   2   :2076
##  4      :1432   Median :2.000   3:1965            3   :1007   3   :4818
##  3      :1106   Mean   :2.658   4:2091            4   : 250   NA's:2151
##  2      : 873   3rd Qu.:4.000   5:2227            5   : 103
##  (Other): 811   Max.   :6.000                     NA's:  33
##  NA's   :2141   NA's   :2096
##    cigdyal_19        BMIVal       NDPNow_19      dnoft_19    drinkYN_19
##  Min.   : 0.000   Min.   : 9.723   1   : 317   4      :1978   1   :1567
##  1st Qu.: 0.000   1st Qu.:21.915   2   :  78   5      :1191   2   :6586
##  Median : 0.000   Median :25.904   3   :  17   3      :1106   NA's:2146
##  Mean   : 1.692   Mean   :26.223   4   :7739   6      : 748
##  3rd Qu.: 0.000   3rd Qu.:29.953   NA's:2148   7      : 705
##  Max.   :60.000   Max.   :73.494               (Other): 977
##  NA's   :2152     NA's   :2224                 NA's   :3594
##    d7many3_19       omsysval          GOR1
##  Min.   :0.000   Min.   : 75.0   8      :1620
##  1st Qu.:0.000   1st Qu.:110.5   2      :1379
##  Median :1.000   Median :121.0   7      :1284
##  Mean   :1.595   Mean   :122.9   6      :1179
##  3rd Qu.:3.000   3rd Qu.:133.5   3      :1138
##  Max.   :7.000   Max.   :209.5   5      : 972
##  NA's   :2147    NA's   :5593    (Other):2727
```

Note that the null flavors may not be used for modeling (and can just be treated as generic missing values), but they will bve useful for evaluating the study design. For example, lots of **Refused** for a variable could mean there is a bias in porivacy or that the question is too sensitive. Lots of **Don't know** for a variable could indicate some recall bias and that the question is poorly designed, whereas lots of **Not applicable** either comes from reduced generalisability (e.g. "Is patient currently pregnant?) or poorly measured variables (Like valid BMI results being sparse due to bad measurements or missing heights/weights).

The variable d7many3_19 has nothing but missing values, so this variable can be dropped from analysis.