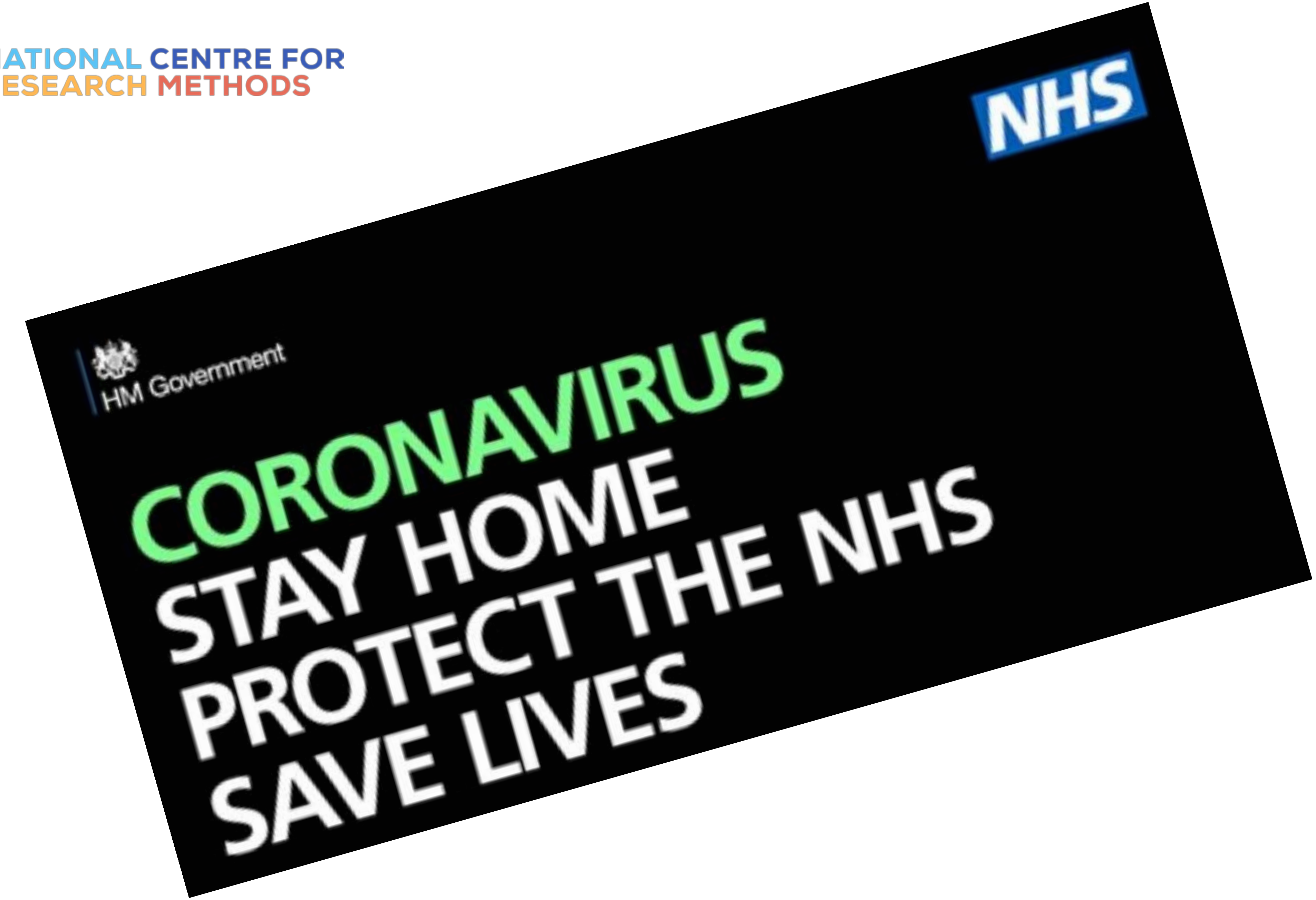
NATIONAL CENTRE FOR RESEARCH METHODS

Reproducible Social Research

Professor Vernon Gayle
vernon.gayle@ed.ac.uk
@Profbigvern
https://github.com/vernongayle

UKRI Economic and Social Research Council
UNIVERSITY OF Southampton
MANCHESTER 1824 The University of Manchester
THE UNIVERSITY OF EDINBURGH


HM Government
NHS
CORONAVIRUS
STAY HOME
PROTECT THE NHS
SAVE LIVES

PRODUCTION
Reproducibility

DIRECTOR
NCRM

DATE
May 20

SCENE
1

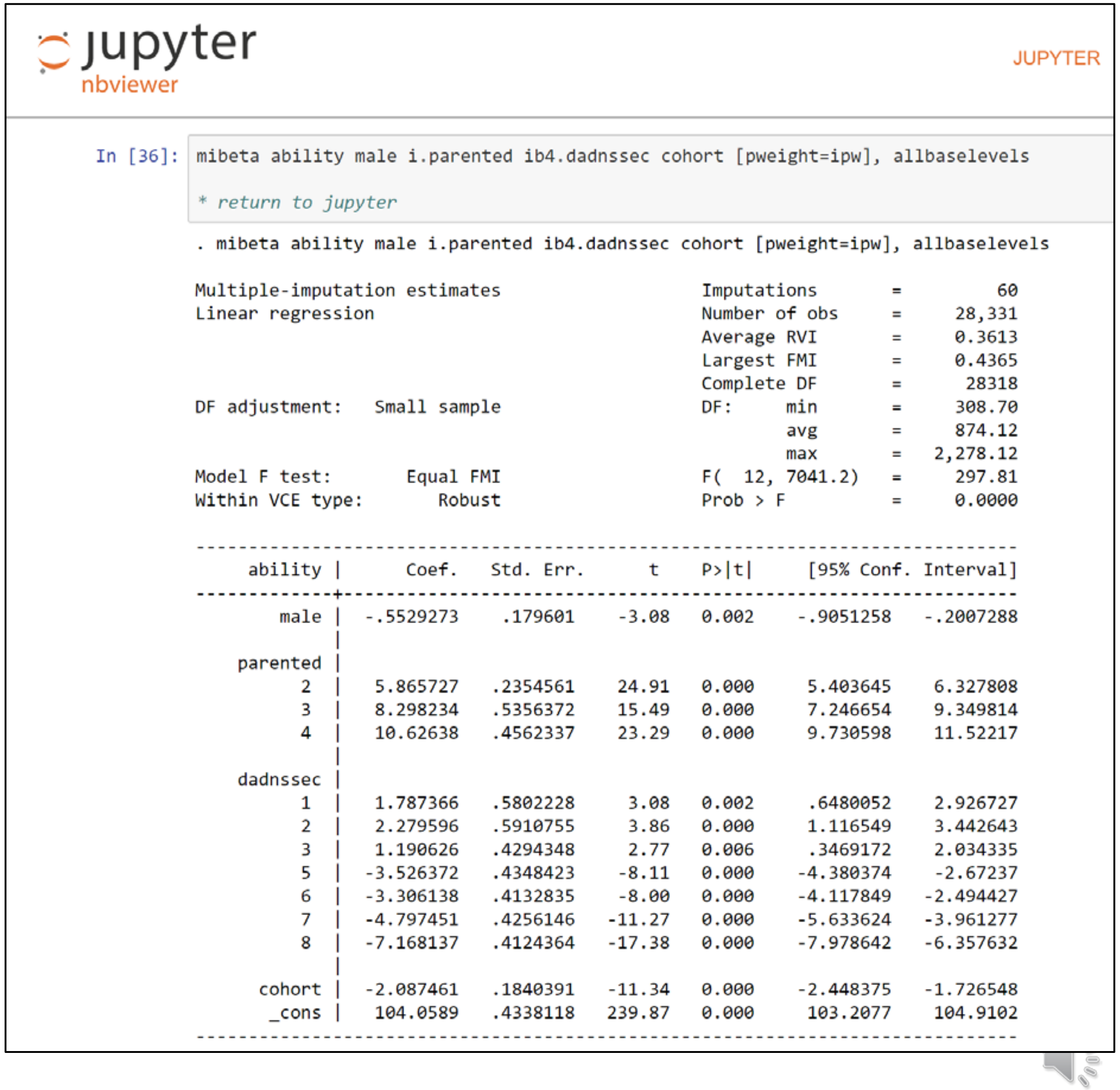
TAKE
100

CAMERA
Vernon Gayle

Traditional Publication

BJS
THE BRITISH JOURNAL OF SOCIOLOGY
Special Issue: Put to the Test - The Sociology of Testing
LSE
WILEY

Additional Material

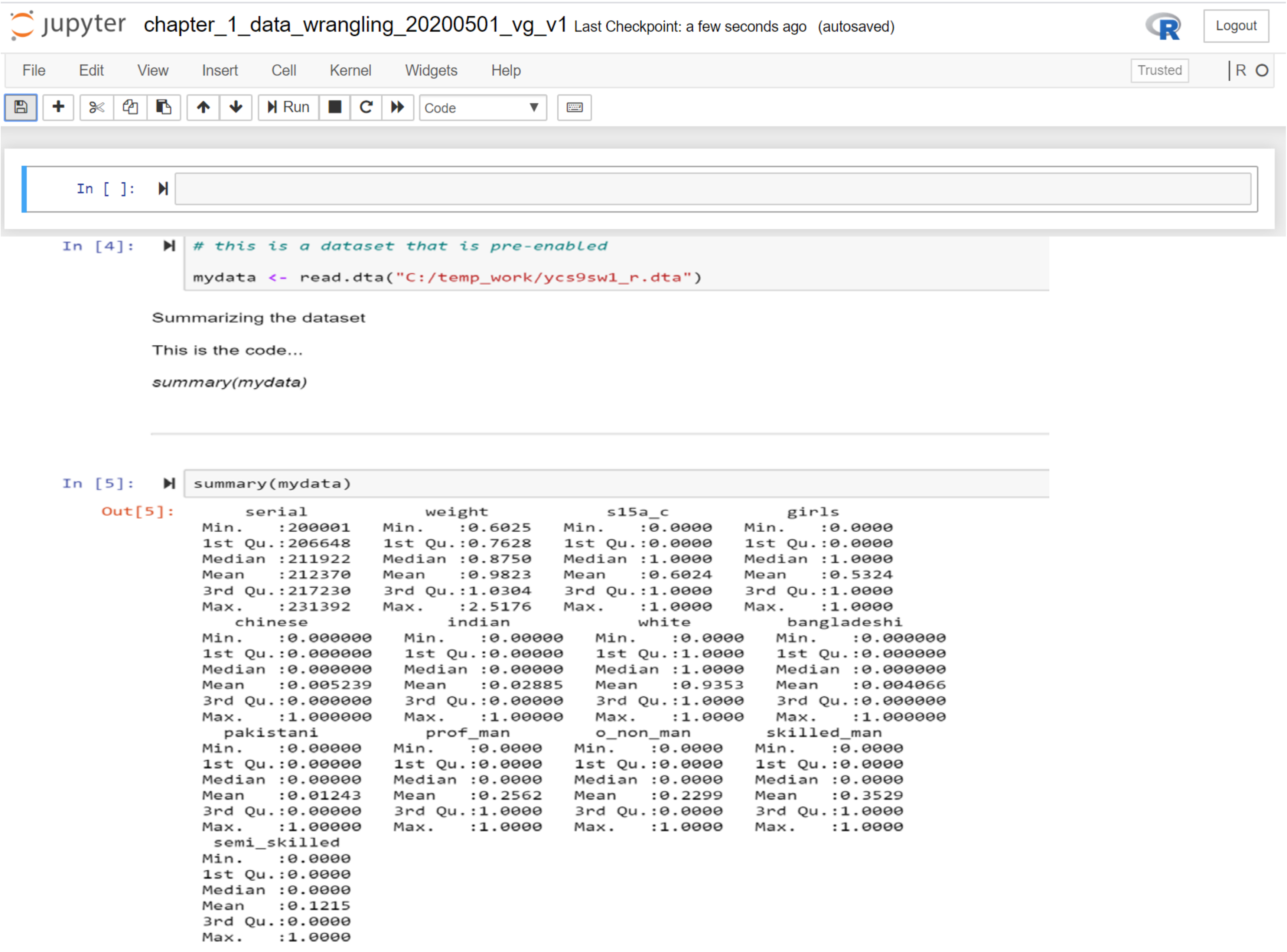


In [36]:

```
mibeta ability male i.parented ib4.dadnssec cohort [pweight=ipw], allbaselevels

* return to jupyter
```

```
. mibeta ability male i.parented ib4.dadnssec cohort [pweight=ipw], allbaselevels

Multiple-imputation estimates                   Imputations       =         60
Linear regression                               Number of obs     =     28,331
                                                Average RVI       =     0.3613
                                                Largest FMI       =     0.4365
                                                Complete DF       =      28318
DF adjustment:   Small sample                   DF:     min       =     308.70
                                                        avg       =     874.12
                                                        max       =   2,278.12
Model F test:       Equal FMI                   F(  12, 7041.2)   =     297.81
Within VCE type:       Robust                   Prob > F          =     0.0000

------------------------------------------------------------------------------
     ability |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.5529273    .179601    -3.08   0.002    -.9051258   -.2007288
             |
    parented |
          2  |   5.865727   .2354561    24.91   0.000     5.403645    6.327808
          3  |   8.298234   .5356372    15.49   0.000     7.246654    9.349814
          4  |   10.62638   .4562337    23.29   0.000     9.730598    11.52217
             |
    dadnssec |
          1  |   1.787366   .5802228     3.08   0.002     .6480052    2.926727
          2  |   2.279596   .5910755     3.86   0.000     1.116549    3.442643
          3  |   1.190626   .4294348     2.77   0.006     .3469172    2.034335
          5  |  -3.526372   .4348423    -8.11   0.000    -4.380374    -2.67237
          6  |  -3.306138   .4132835    -8.00   0.000    -4.117849   -2.494427
          7  |  -4.797451   .4256146   -11.27   0.000    -5.633624   -3.961277
          8  |  -7.168137   .4124364   -17.38   0.000    -7.978642   -6.357632
             |
      cohort |  -2.087461   .1840391   -11.34   0.000    -2.448375   -1.726548
       _cons |   104.0589   .4338118   239.87   0.000     103.2077    104.9102
------------------------------------------------------------------------------
```

The Problem


jupyter
chapter_1_data_wrangling_20200501_vg_v1 Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
R
Run
Code

```
In [ ]:
```

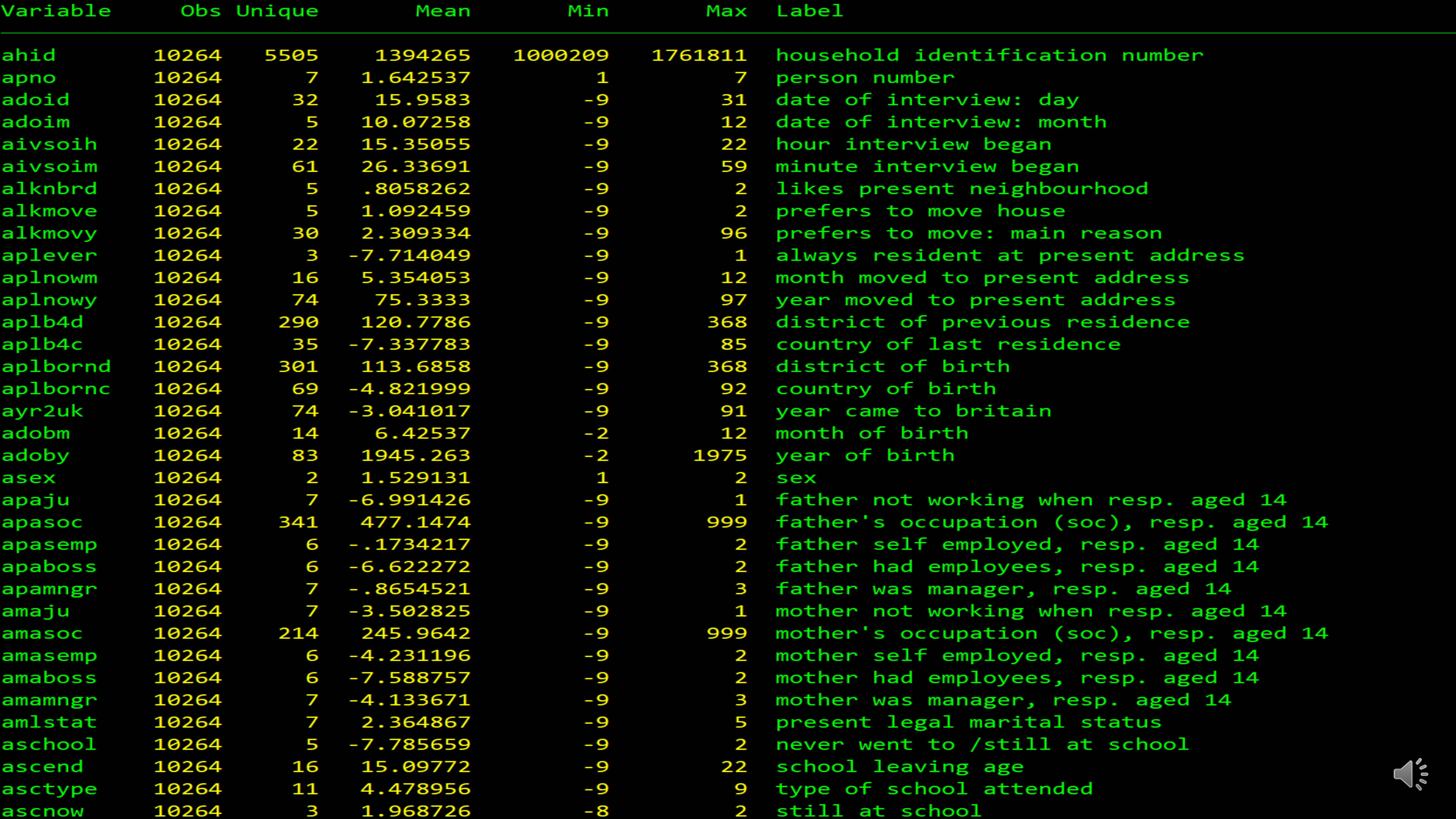
```
In [4]:  # this is a dataset that is pre-enabled
         mydata <- read.dta("C:/temp_work/ycs9sw1_r.dta")
```

Summarizing the dataset

This is the code...

summary(mydata)

```
In [5]:  summary(mydata)

Out[5]:
     serial           weight           s15a_c           girls
 Min.   :200001   Min.   :0.6025   Min.   :0.0000   Min.   :0.0000
 1st Qu.:206648   1st Qu.:0.7628   1st Qu.:0.0000   1st Qu.:0.0000
 Median :211922   Median :0.8750   Median :1.0000   Median :1.0000
 Mean   :212370   Mean   :0.9823   Mean   :0.6024   Mean   :0.5324
 3rd Qu.:217230   3rd Qu.:1.0304   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :231392   Max.   :2.5176   Max.   :1.0000   Max.   :1.0000
    chinese              indian              white         bangladeshi
 Min.   :0.000000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000000
 1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.000000
 Median :0.000000   Median :0.00000   Median :1.0000   Median :0.000000
 Mean   :0.005239   Mean   :0.02885   Mean   :0.9353   Mean   :0.004066
 3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.000000
 Max.   :1.000000   Max.   :1.00000   Max.   :1.0000   Max.   :1.000000
   pakistani          prof_man         o_non_man       skilled_man
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.01243   Mean   :0.2562   Mean   :0.2299   Mean   :0.3529
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  semi_skilled
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1215
 3rd Qu.:0.0000
 Max.   :1.0000
```



```
Variable        Obs  Unique        Mean        Min        Max  Label
------------------------------------------------------------------------------------------------------
ahid          10264    5505     1394265    1000209    1761811  household identification number
apno          10264       7    1.642537          1          7  person number
adoid         10264      32     15.9583         -9         31  date of interview: day
adoim         10264       5    10.07258         -9         12  date of interview: month
aivsoih       10264      22    15.35055         -9         22  hour interview began
aivsoim       10264      61    26.33691         -9         59  minute interview began
alknbrd       10264       5    .8058262         -9          2  likes present neighbourhood
alkmove       10264       5    1.092459         -9          2  prefers to move house
alkmovy       10264      30    2.309334         -9         96  prefers to move: main reason
aplever       10264       3   -7.714049         -9          1  always resident at present address
aplnowm       10264      16    5.354053         -9         12  month moved to present address
aplnowy       10264      74     75.3333         -9         97  year moved to present address
aplb4d        10264     290    120.7786         -9        368  district of previous residence
aplb4c        10264      35   -7.337783         -9         85  country of last residence
aplbornd      10264     301    113.6858         -9        368  district of birth
aplbornc      10264      69   -4.821999         -9         92  country of birth
ayr2uk        10264      74   -3.041017         -9         91  year came to britain
adobm         10264      14     6.42537         -2         12  month of birth
adoby         10264      83    1945.263         -2       1975  year of birth
asex          10264       2    1.529131          1          2  sex
apaju         10264       7   -6.991426         -9          1  father not working when resp. aged 14
apasoc        10264     341    477.1474         -9        999  father's occupation (soc), resp. aged 14
apasemp       10264       6   -.1734217         -9          2  father self employed, resp. aged 14
apaboss       10264       6   -6.622272         -9          2  father had employees, resp. aged 14
apamngr       10264       7   -.8654521         -9          3  father was manager, resp. aged 14
amaju         10264       7   -3.502825         -9          1  mother not working when resp. aged 14
amasoc        10264     214    245.9642         -9        999  mother's occupation (soc), resp. aged 14
amasemp       10264       6   -4.231196         -9          2  mother self employed, resp. aged 14
amaboss       10264       6   -7.588757         -9          2  mother had employees, resp. aged 14
amamngr       10264       7   -4.133671         -9          3  mother was manager, resp. aged 14
amlstat       10264       7    2.364867         -9          5  present legal marital status
aschool       10264       5   -7.785659         -9          2  never went to /still at school
ascend        10264      16    15.09772         -9         22  school leaving age
asctype       10264      11    4.478956         -9          9  type of school attended
ascnow        10264       3    1.968726         -8          2  still at school
```

https://www.alfadispenser.com/2018/05/30/repeatability-vs-reproducibility-a-little-talk/

The Case for Greater Transparency

Greater transparency will

1. Increase the capacity to understand how the research was conducted
2. Help other scholars evaluate the analyses undertaken
3. Aid the detection of errors and inconsistencies
4. Facilitate the incremental development of work
5. Contribute to limiting negative research practices
6. Provide extra safeguards against nefarious practices
7. Improve confidence in results within and beyond the academic community

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Duplication and Replication

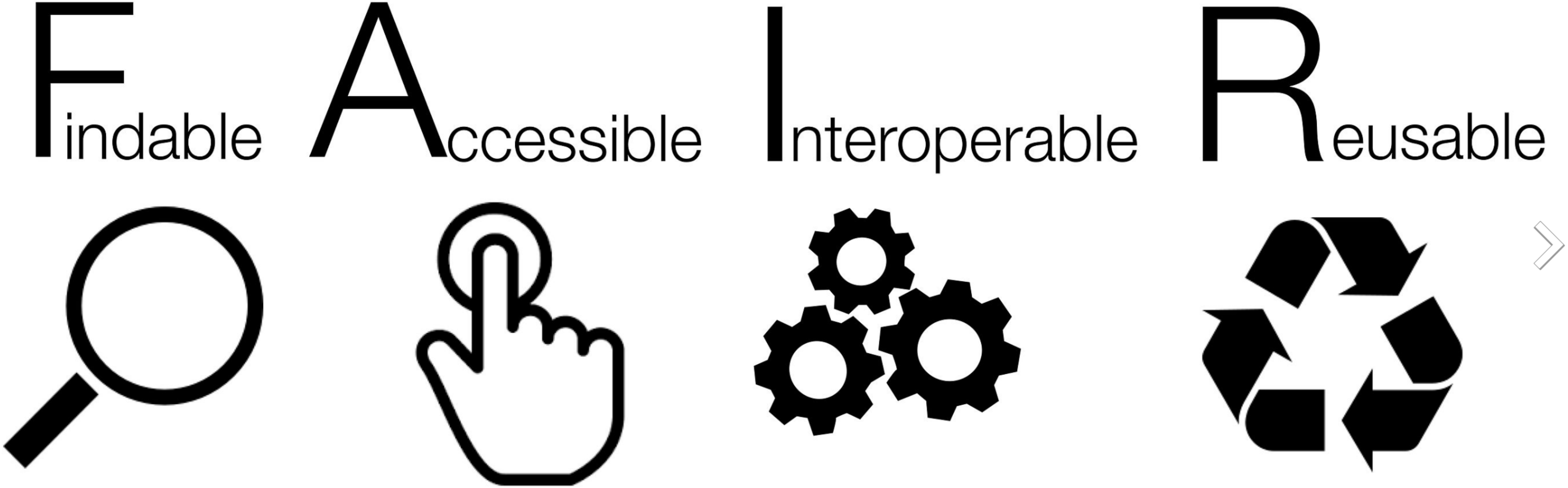
A replication study extends the original work with

1. additional measures
2. alternative measures
3. new data
4. different statistical analytical techniques

or any combination of these four components

NATIONAL CENTRE FOR RESEARCH METHODS

Data Sharing and Citing Data

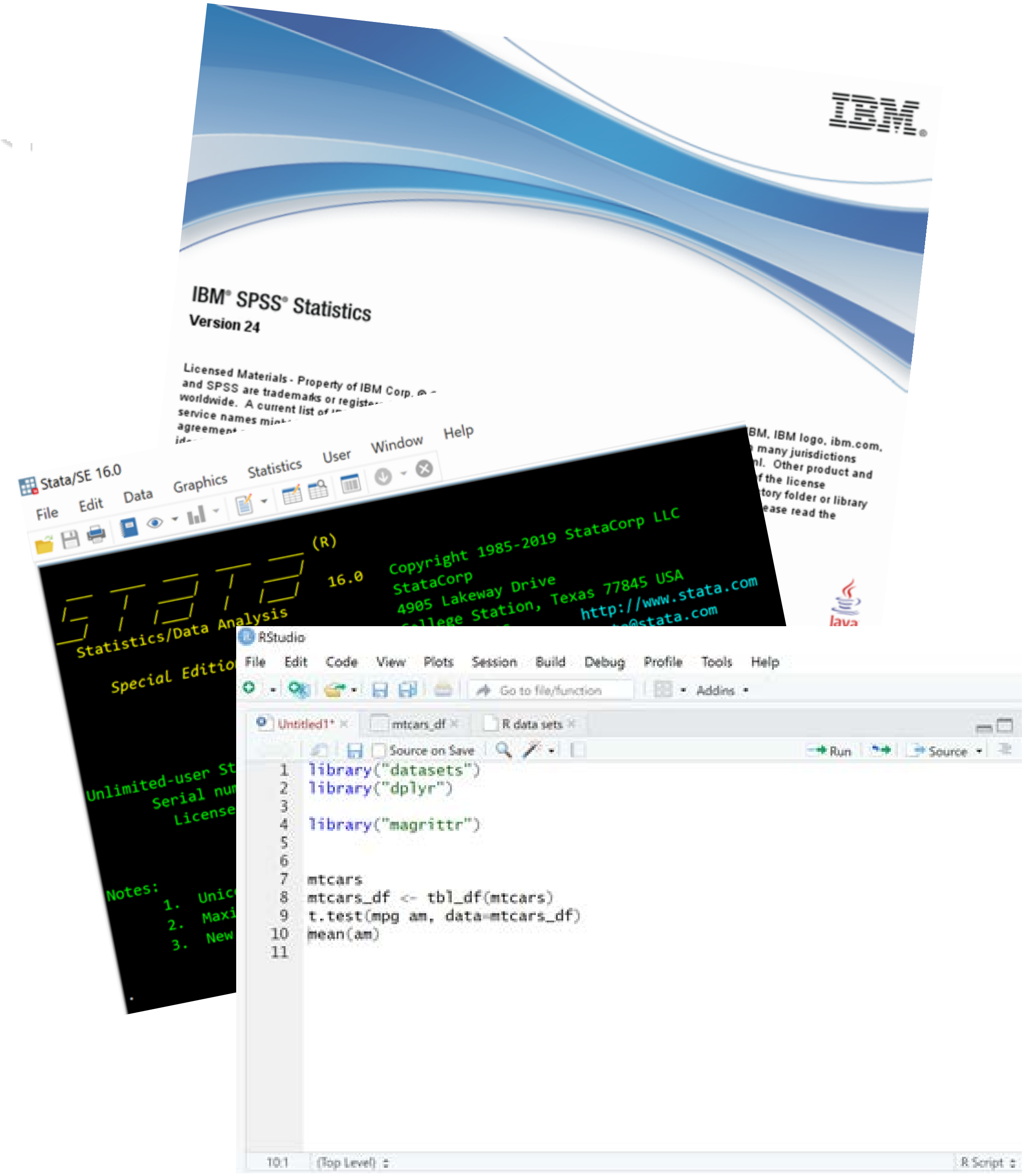


https://en.wikipedia.org/wiki/FAIR_data#/media/File:FAIR_data_principles.jpg

For Your Eyes Only

The Workflow and Code Sharing

OXFORD SCIENCE PUBLICATIONS
NAG
GLIM 4
THE STATISTICAL SYSTEM FOR GENERALIZED LINEAR INTERACTIVE MODELLING
Placebo
Treatment



IBM® SPSS® Statistics
Version 24

Licensed Materials - Property of IBM Corp. ©
and SPSS are trademarks or registe
worldwide. A current list of
service names mig
agreement
id

BM, IBM logo, ibm.com,
many jurisdictions
nl. Other product and
f the license
tory folder or library
ease read the

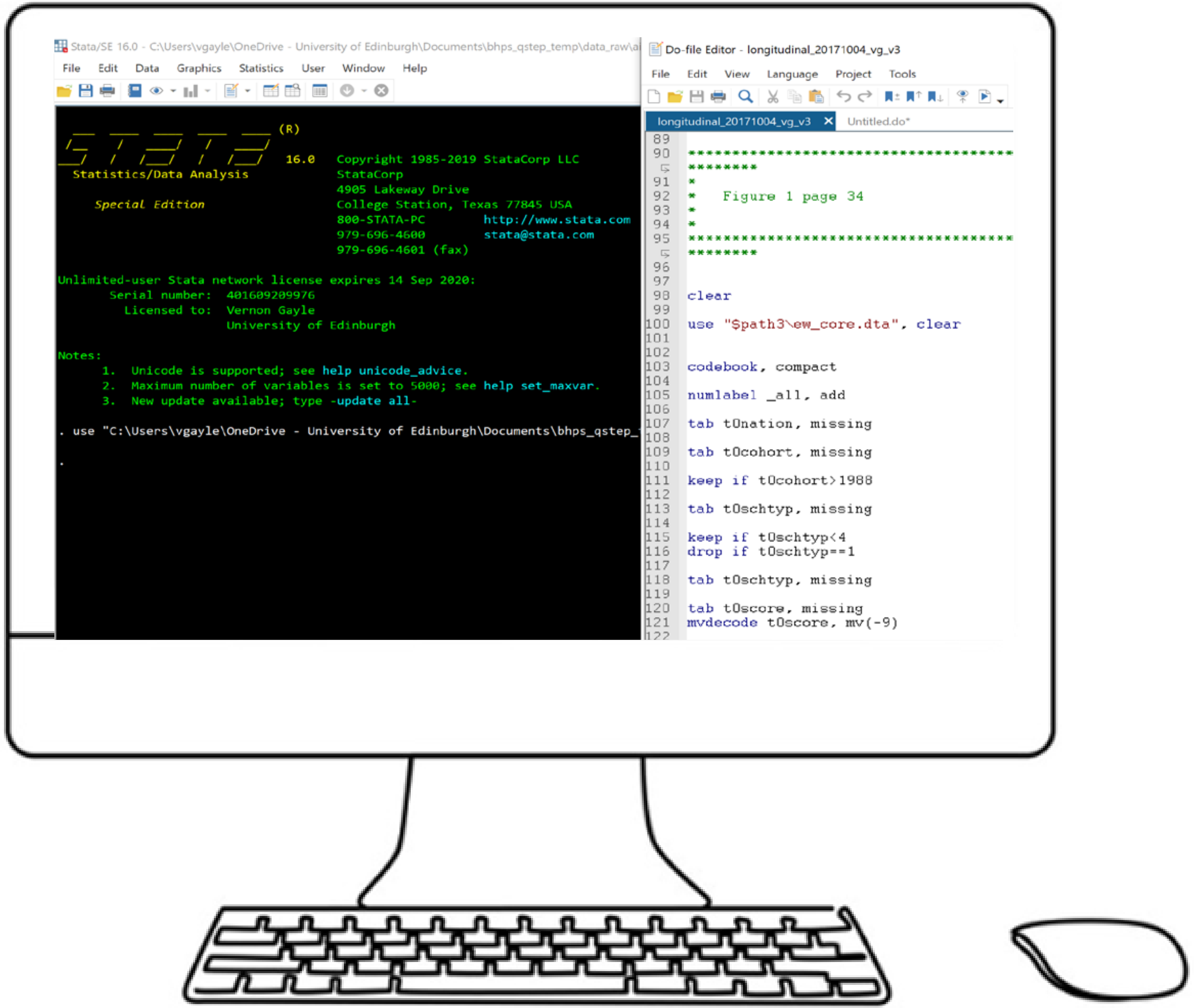
```
                     (R)
 STATA        16.0   Copyright 1985-2019 StataCorp LLC
 Statistics/Data Analysis    StataCorp
                     4905 Lakeway Drive
   Special Editio    College Station, Texas 77845 USA
                     http://www.stata.com
                     @stata.com

 Unlimited-user St
       Serial num
         License

 Notes:
       1.  Unic
       2.  Maxi
       3.  New
```

```
library("datasets")
library("dplyr")

library("magrittr")


mtcars
mtcars_df <- tbl_df(mtcars)
t.test(mpg am, data=mtcars_df)
mean(am)

```

Drop down menus = no audit trail

GUIs will leave you in a sticky mess!

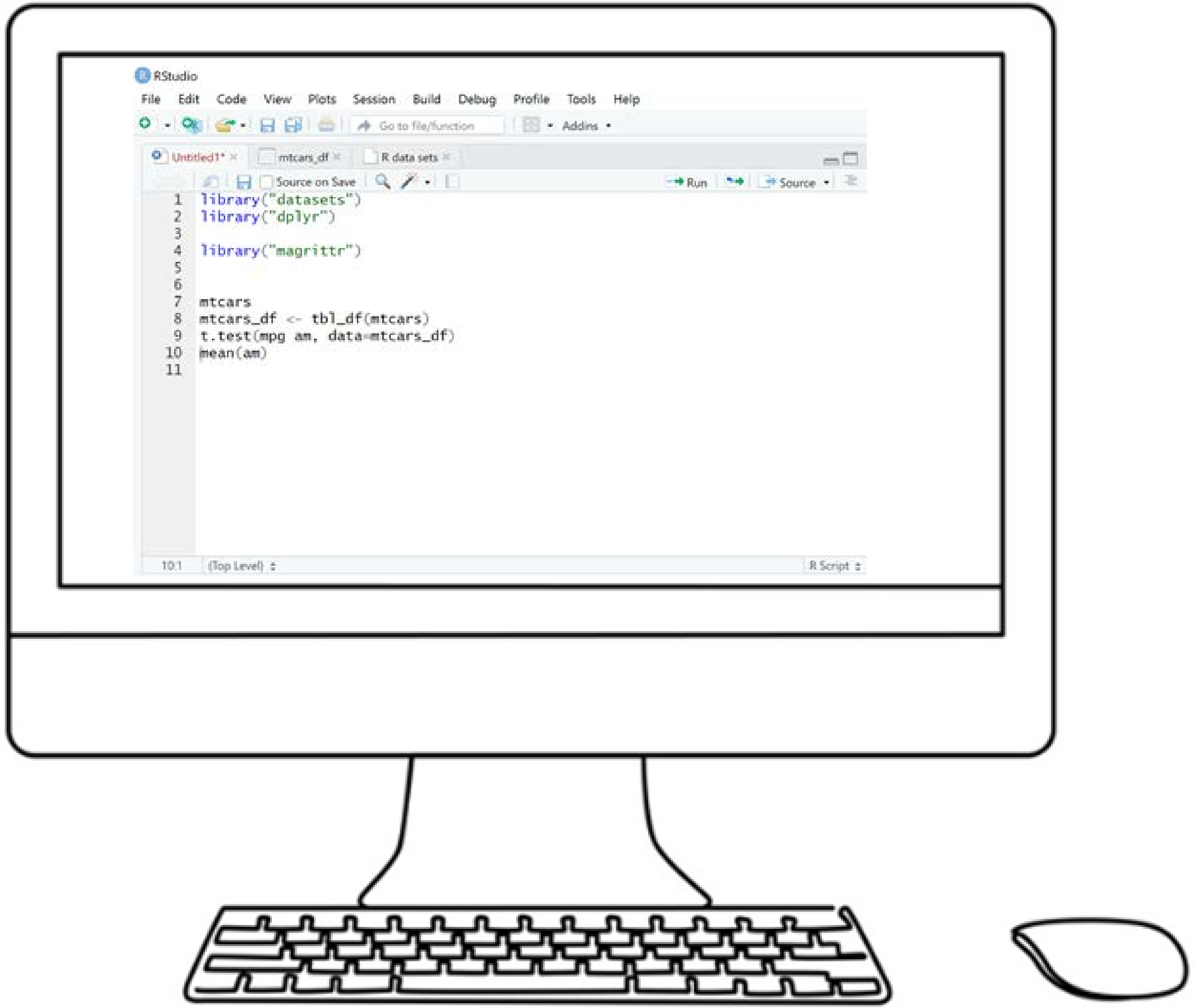
```
get stata file = 'd:datadata13.dta'.
```



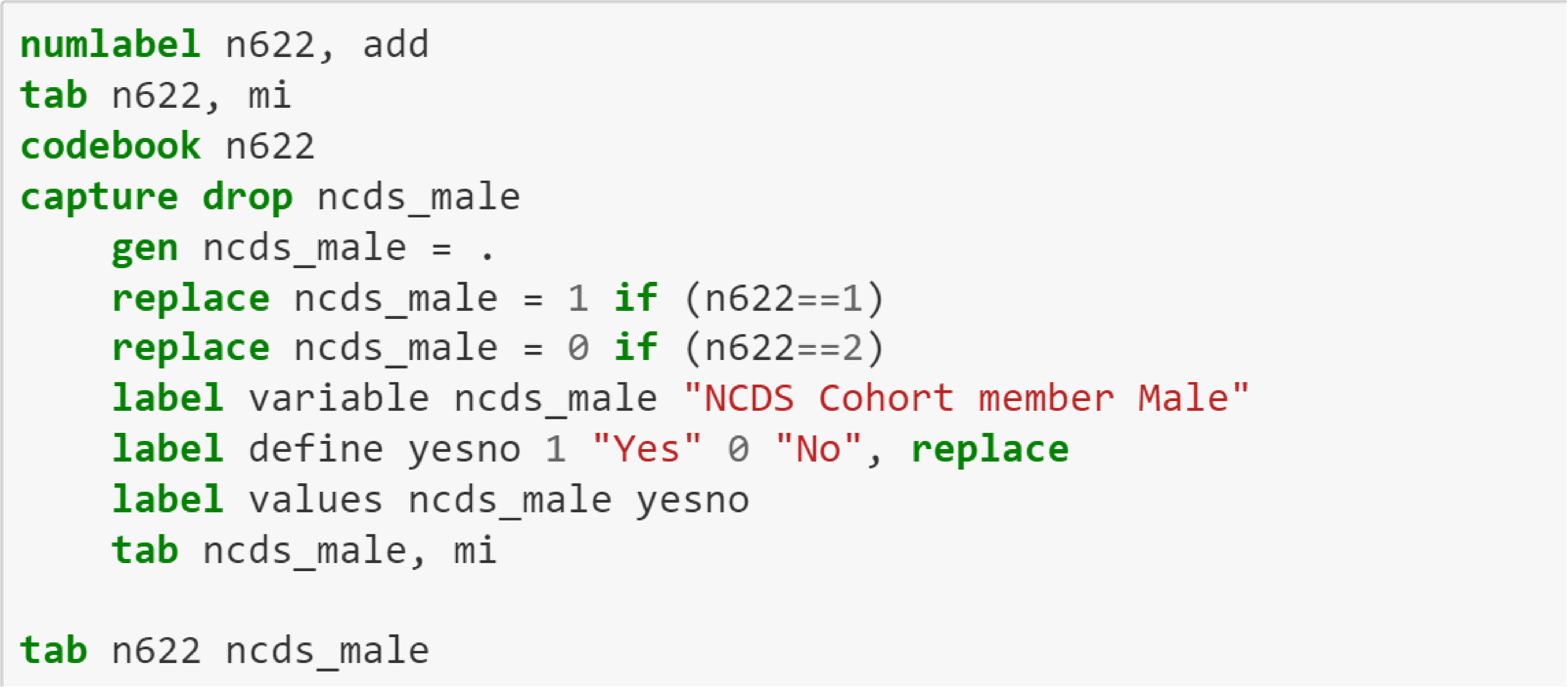
Stata/SE 16.0 - C:\Users\vgayle\OneDrive - University of Edinburgh\Documents\bhps_qstep_temp\data_raw\a
File Edit Data Graphics Statistics User Window Help
Do-file Editor - longitudinal_20171004_vg_v3
File Edit View Language Project Tools
longitudinal_20171004_vg_v3
Untitled.do*

```
                                          (R)
                                     16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis                  StataCorp
                                            4905 Lakeway Drive
     Special Edition                        College Station, Texas 77845 USA
                                            800-STATA-PC        http://www.stata.com
                                            979-696-4600        stata@stata.com
                                            979-696-4601 (fax)

Unlimited-user Stata network license expires 14 Sep 2020:
       Serial number:  401609209976
         Licensed to:  Vernon Gayle
                       University of Edinburgh

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

. use "C:\Users\vgayle\OneDrive - University of Edinburgh\Documents\bhps_qstep_

.
```

```

*************************************
********
*
*    Figure 1 page 34
*
*
*************************************
********


clear

use "$path3\ew_core.dta", clear


codebook, compact

numlabel _all, add

tab t0nation, missing

tab t0cohort, missing

keep if t0cohort>1988

tab t0schtyp, missing

keep if t0schtyp<4
drop if t0schtyp==1

tab t0schtyp, missing

tab t0score, missing
mvdecode t0score, mv(-9)
```

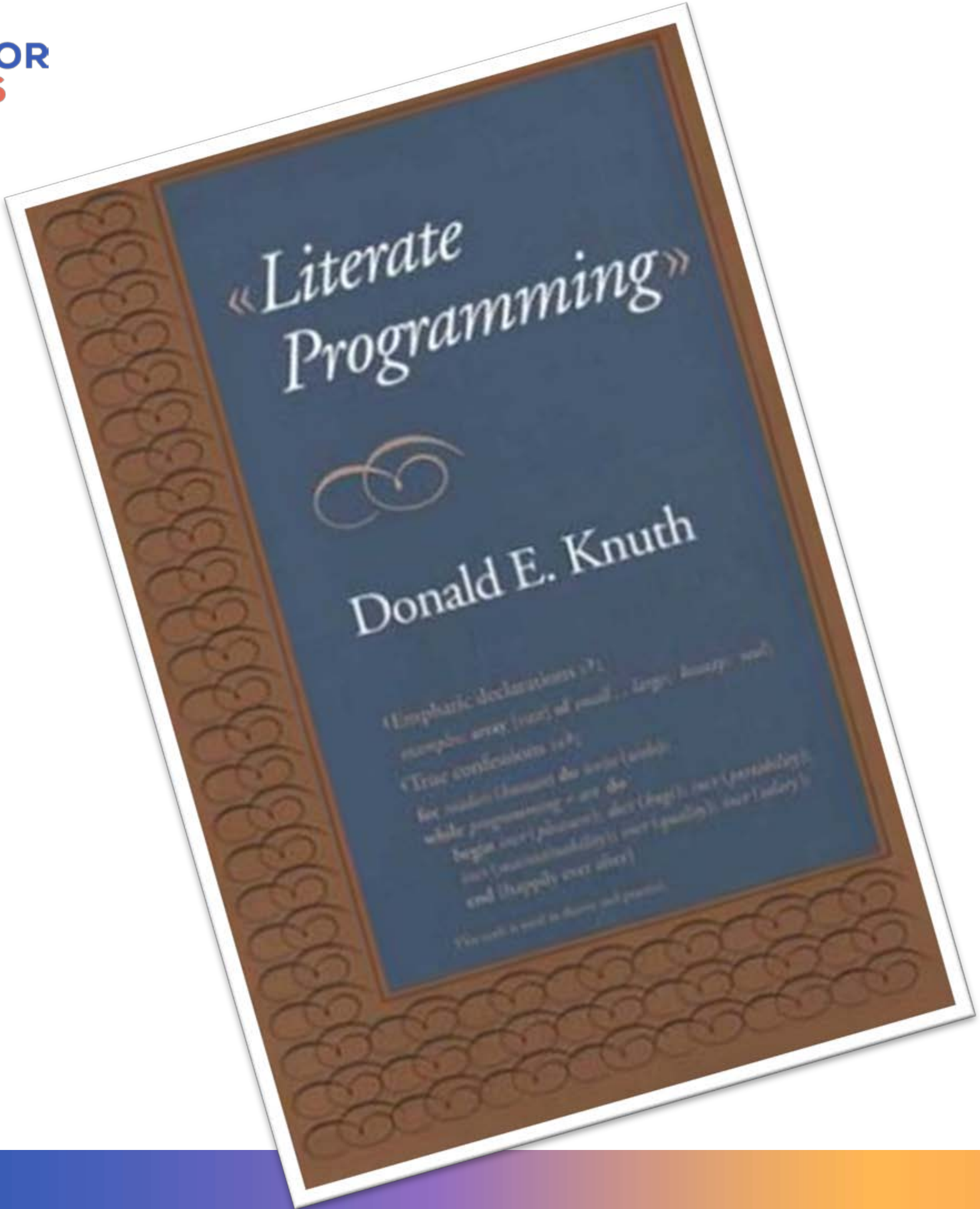


```
library("datasets")
library("dplyr")

library("magrittr")


mtcars
mtcars_df <- tbl_df(mtcars)
t.test(mpg am, data=mtcars_df)
mean(am)

```



```
numlabel n622, add
tab n622, mi
codebook n622
capture drop ncds_male
    gen ncds_male = .
    replace ncds_male = 1 if (n622==1)
    replace ncds_male = 0 if (n622==2)
    label variable ncds_male "NCDS Cohort member Male"
    label define yesno 1 "Yes" 0 "No", replace
    label values ncds_male yesno
    tab ncds_male, mi

tab n622 ncds_male
```

Documenting the Workflow



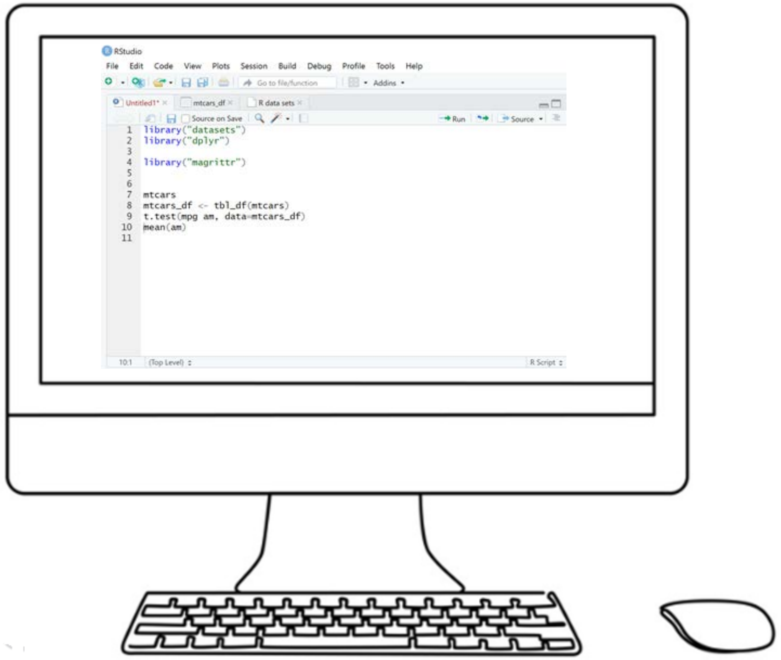
«Literate Programming»

Donald E. Knuth

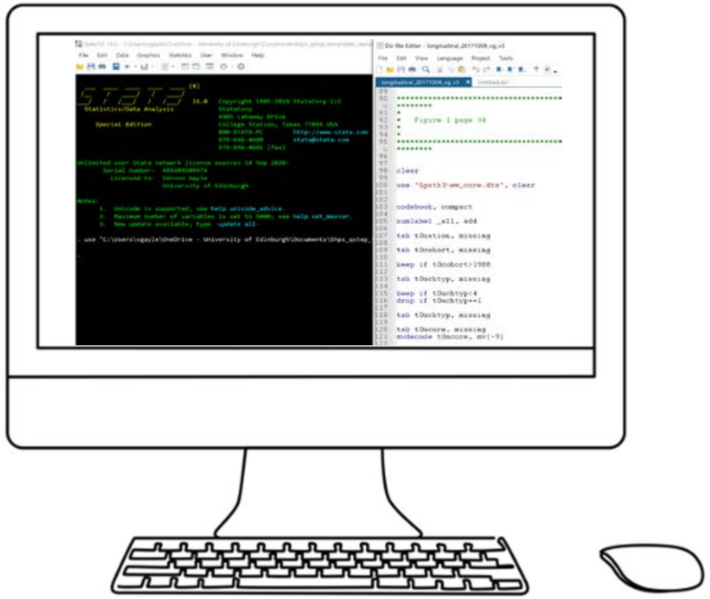
⟨Emphatic declarations 17⟩;
[illegible] array [illegible] of small . . large; [illegible]
⟨True confessions 14⟩;
for [illegible] the [illegible]
while programming = are the
begin [illegible]; [illegible]; [illegible]
[illegible]; [illegible]; [illegible]
end (happily ever after)

This work is [illegible] theory and practice.

```
get stata file = 'd:datadata13.dta'.
```

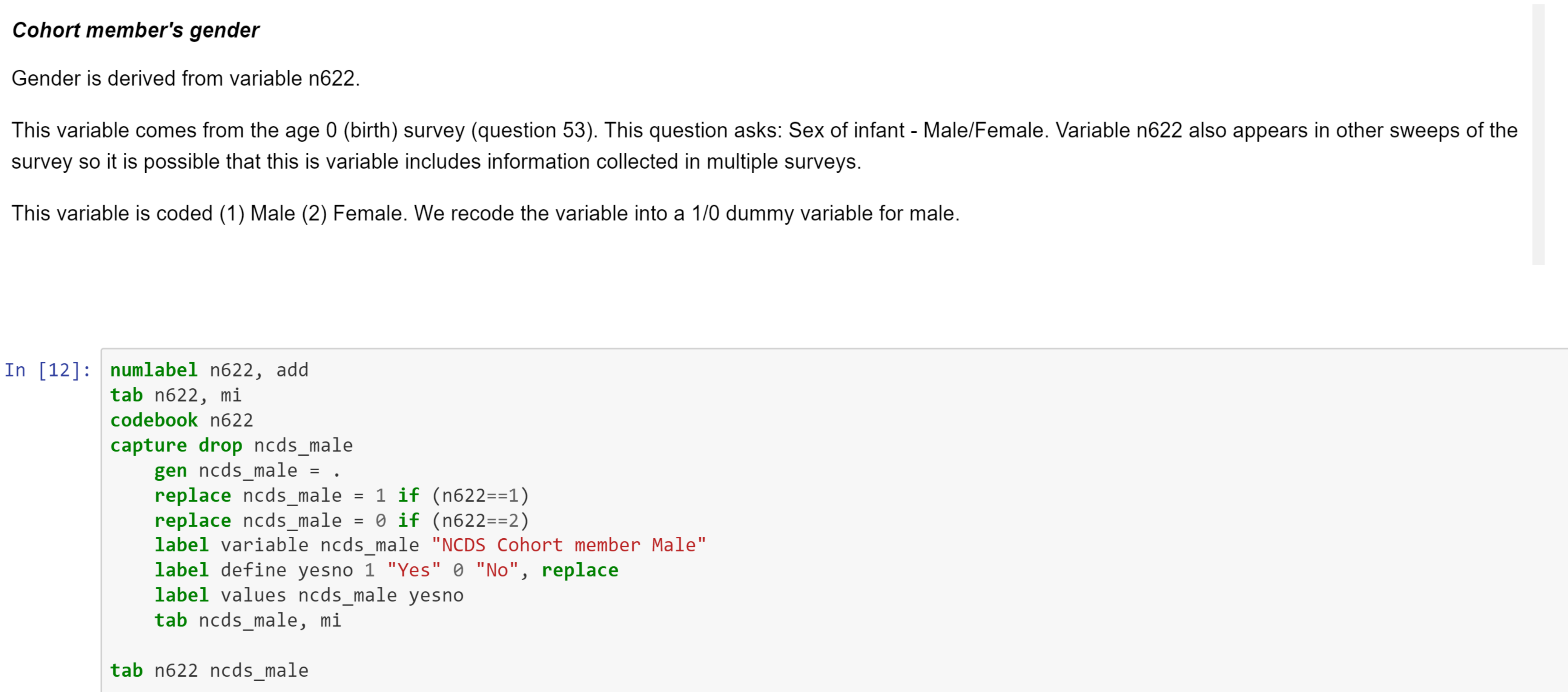


```
library("datasets")
library("dplyr")

library("magrittr")


mtcars
mtcars_df <- tbl_df(mtcars)
t.test(mpg am, data=mtcars_df)
mean(am)
```


Stata/SE 16.0 - C:\Users\vgayle\OneDrive - University of Edinburgh\Documents\bhps_qstep_temp\stata_revis
File Edit Data Graphics Statistics User Window Help
Do-file Editor - longitudinal_20171004_vg_v3
File Edit View Language Project Tools
longitudinal_20171004_vg_v3
Untitled.do*

```
  STATA (R)
                          16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis       StataCorp
                                 4905 Lakeway Drive
     Special Edition             College Station, Texas 77845 USA
                                 800-STATA-PC        http://www.stata.com
                                 979-696-4600        stata@stata.com
                                 979-696-4601 (fax)

Unlimited-user Stata network license expires 18 Sep 2020:
       Serial number:  [illegible]
         Licensed to:  Vernon Gayle
                       University of Edinburgh

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

. use "C:\Users\vgayle\OneDrive - University of Edinburgh\Documents\bhps_qstep_

.
```

```
*******************************************
********
*
*   Figure 1 page 34
*
*
*******************************************
********


clear

use "$path3\wm_core.dta", clear


codebook, compact

numlabel _all, add

tab t0ustion, missing

tab t0cohort, missing

keep if t0cohort>1988

tab t0schtyp, missing

keep if t0schtyp<4
drop if t0schtyp==1

tab t0schtyp, missing

tab t0score, missing
mvdecode t0score, mv(-9)
```



Cohort member's gender

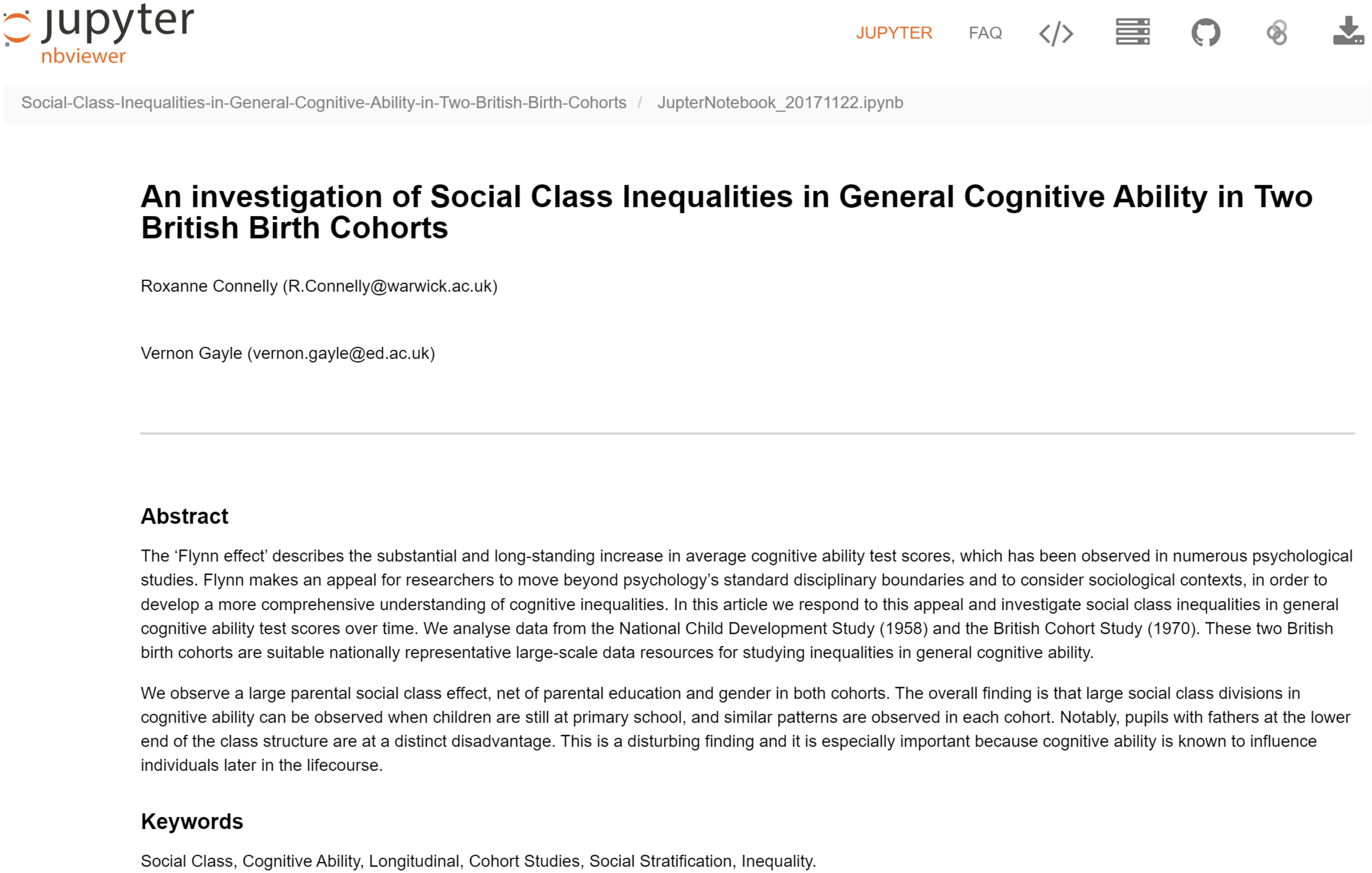
Gender is derived from variable n622.

This variable comes from the age 0 (birth) survey (question 53). This question asks: Sex of infant - Male/Female. Variable n622 also appears in other sweeps of the survey so it is possible that this is variable includes information collected in multiple surveys.

This variable is coded (1) Male (2) Female. We recode the variable into a 1/0 dummy variable for male.

In [12]:

```
numlabel n622, add
tab n622, mi
codebook n622
capture drop ncds_male
    gen ncds_male = .
    replace ncds_male = 1 if (n622==1)
    replace ncds_male = 0 if (n622==2)
    label variable ncds_male "NCDS Cohort member Male"
    label define yesno 1 "Yes" 0 "No", replace
    label values ncds_male yesno
    tab ncds_male, mi

tab n622 ncds_male
```


jupyter
nbviewer
JUPYTER
FAQ
Social-Class-Inequalities-in-General-Cognitive-Ability-in-Two-British-Birth-Cohorts / JupterNotebook_20171122.ipynb

An investigation of Social Class Inequalities in General Cognitive Ability in Two British Birth Cohorts

Roxanne Connelly (R.Connelly@warwick.ac.uk)

Vernon Gayle (vernon.gayle@ed.ac.uk)

Abstract

The ‘Flynn effect’ describes the substantial and long-standing increase in average cognitive ability test scores, which has been observed in numerous psychological studies. Flynn makes an appeal for researchers to move beyond psychology’s standard disciplinary boundaries and to consider sociological contexts, in order to develop a more comprehensive understanding of cognitive inequalities. In this article we respond to this appeal and investigate social class inequalities in general cognitive ability test scores over time. We analyse data from the National Child Development Study (1958) and the British Cohort Study (1970). These two British birth cohorts are suitable nationally representative large-scale data resources for studying inequalities in general cognitive ability.

We observe a large parental social class effect, net of parental education and gender in both cohorts. The overall finding is that large social class divisions in cognitive ability can be observed when children are still at primary school, and similar patterns are observed in each cohort. Notably, pupils with fathers at the lower end of the class structure are at a distinct disadvantage. This is a disturbing finding and it is especially important because cognitive ability is known to influence individuals later in the lifecourse.

Keywords

Social Class, Cognitive Ability, Longitudinal, Cohort Studies, Social Stratification, Inequality.

Cohort member's gender

Gender is derived from variable n622.

This variable comes from the age 0 (birth) survey (question 53). This question asks: Sex of infant - Male/Female. Variable n622 also appears in other sweeps of the survey so it is possible that this is variable includes information collected in multiple surveys.

This variable is coded (1) Male (2) Female. We recode the variable into a 1/0 dummy variable for male.

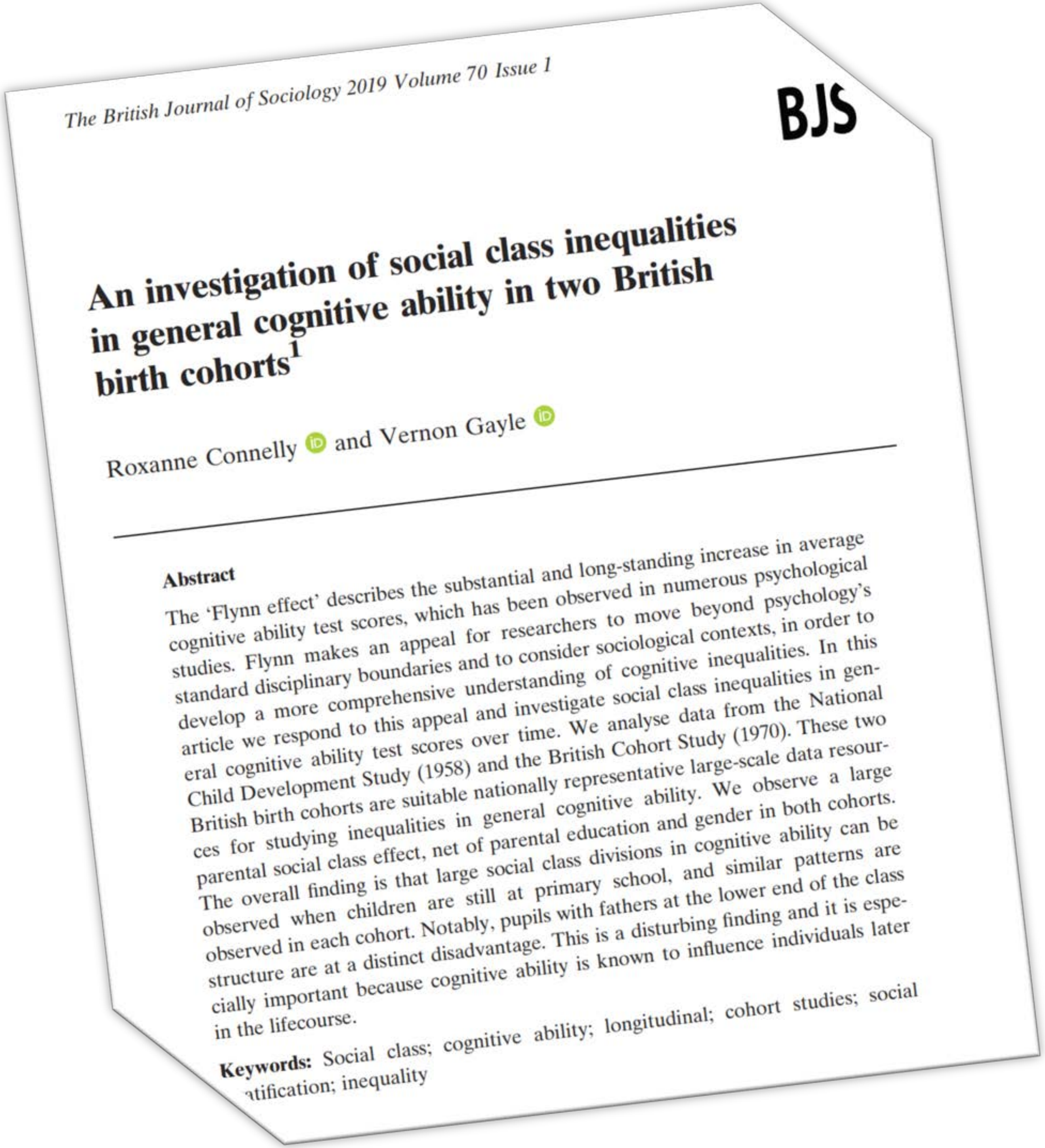
In [12]:

```
numlabel n622, add
tab n622, mi
codebook n622
capture drop ncds_male
    gen ncds_male = .
    replace ncds_male = 1 if (n622==1)
    replace ncds_male = 0 if (n622==2)
    label variable ncds_male "NCDS Cohort member Male"
    label define yesno 1 "Yes" 0 "No", replace
    label values ncds_male yesno
    tab ncds_male, mi

tab n622 ncds_male

*return to jupyter
```

```
. numlabel n622, add

. tab n622, mi

  0-3D Sex of |
        child |      Freq.     Percent        Cum.
--------------+-----------------------------------
      1. Male |      9,595       51.70       51.70
    2. Female |      8,959       48.28       99.98
            . |          4        0.02      100.00
--------------+-----------------------------------
        Total |     18,558      100.00

. codebook n622

-------------------------------------------------------------------------------------------------------------------------------
```

Making the Workflow Public

Current Good Practice



The British Journal of Sociology 2019 Volume 70 Issue 1

BJS

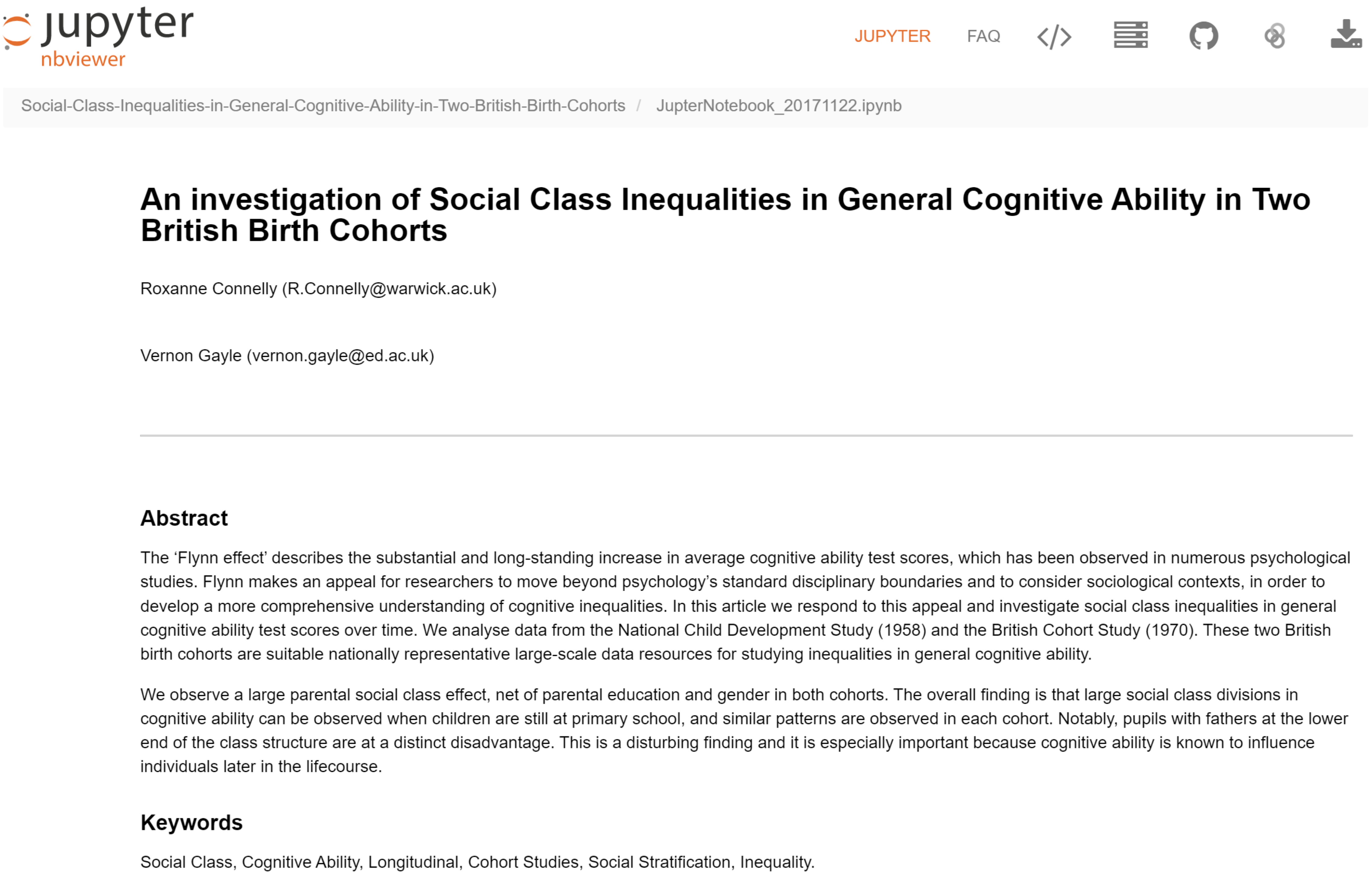
An investigation of social class inequalities in general cognitive ability in two British birth cohorts[1]

Roxanne Connelly and Vernon Gayle

Abstract

The 'Flynn effect' describes the substantial and long-standing increase in average cognitive ability test scores, which has been observed in numerous psychological studies. Flynn makes an appeal for researchers to move beyond psychology's standard disciplinary boundaries and to consider sociological contexts, in order to develop a more comprehensive understanding of cognitive inequalities. In this article we respond to this appeal and investigate social class inequalities in general cognitive ability test scores over time. We analyse data from the National Child Development Study (1958) and the British Cohort Study (1970). These two British birth cohorts are suitable nationally representative large-scale data resources for studying inequalities in general cognitive ability. We observe a large parental social class effect, net of parental education and gender in both cohorts. The overall finding is that large social class divisions in cognitive ability can be observed when children are still at primary school, and similar patterns are observed in each cohort. Notably, pupils with fathers at the lower end of the class structure are at a distinct disadvantage. This is a disturbing finding and it is especially important because cognitive ability is known to influence individuals later in the lifecourse.

Keywords: Social class; cognitive ability; longitudinal; cohort studies; social atification; inequality



jupyter
nbviewer
JUPYTER
FAQ
Social-Class-Inequalities-in-General-Cognitive-Ability-in-Two-British-Birth-Cohorts / JupterNotebook_20171122.ipynb

An investigation of Social Class Inequalities in General Cognitive Ability in Two British Birth Cohorts

Roxanne Connelly (R.Connelly@warwick.ac.uk)

Vernon Gayle (vernon.gayle@ed.ac.uk)

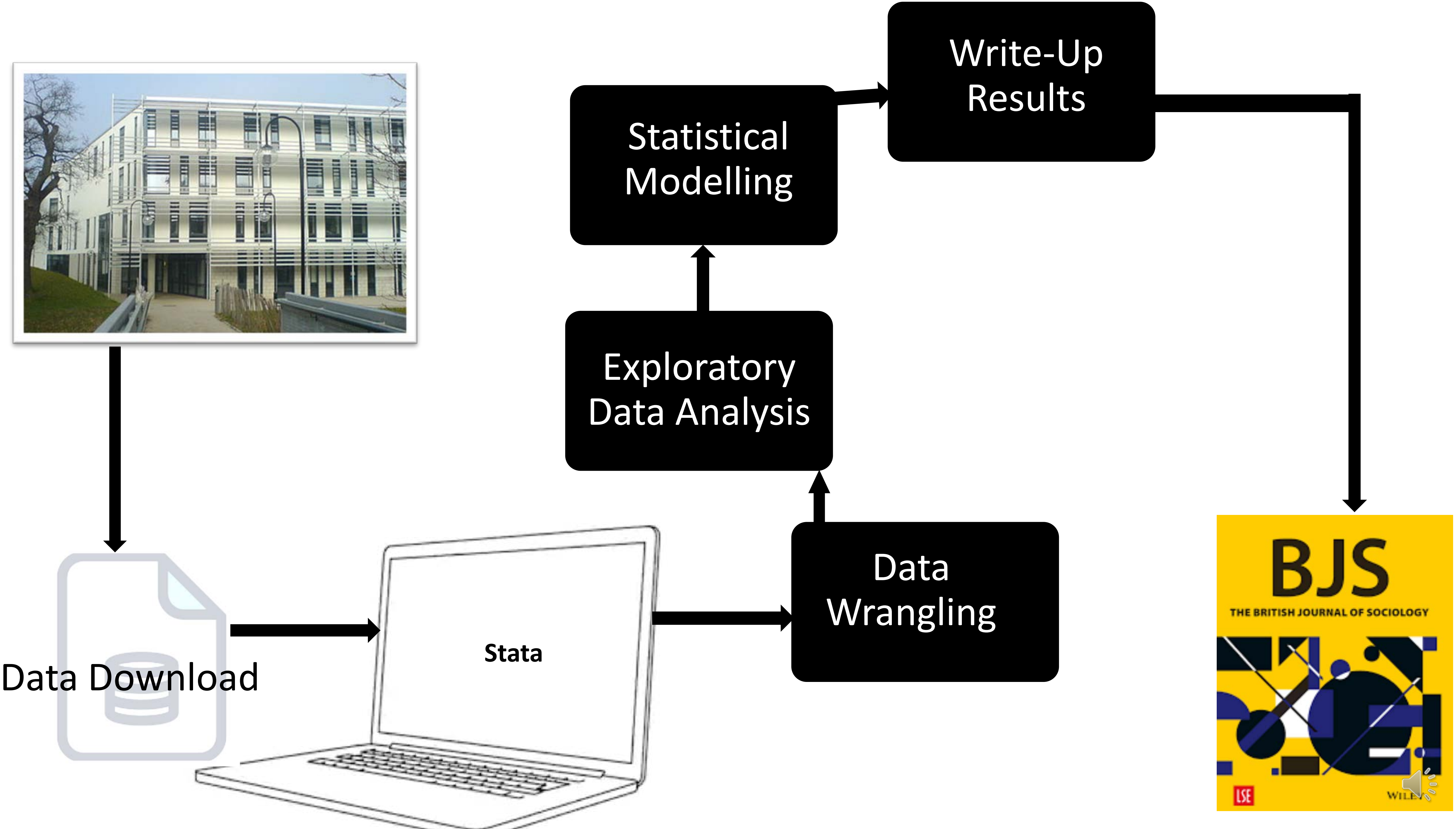
Abstract

The ‘Flynn effect’ describes the substantial and long-standing increase in average cognitive ability test scores, which has been observed in numerous psychological studies. Flynn makes an appeal for researchers to move beyond psychology’s standard disciplinary boundaries and to consider sociological contexts, in order to develop a more comprehensive understanding of cognitive inequalities. In this article we respond to this appeal and investigate social class inequalities in general cognitive ability test scores over time. We analyse data from the National Child Development Study (1958) and the British Cohort Study (1970). These two British birth cohorts are suitable nationally representative large-scale data resources for studying inequalities in general cognitive ability.

We observe a large parental social class effect, net of parental education and gender in both cohorts. The overall finding is that large social class divisions in cognitive ability can be observed when children are still at primary school, and similar patterns are observed in each cohort. Notably, pupils with fathers at the lower end of the class structure are at a distinct disadvantage. This is a disturbing finding and it is especially important because cognitive ability is known to influence individuals later in the lifecourse.

Keywords

Social Class, Cognitive Ability, Longitudinal, Cohort Studies, Social Stratification, Inequality.



Data Download

Stata

Data Wrangling

Exploratory Data Analysis

Statistical Modelling

Write-Up Results

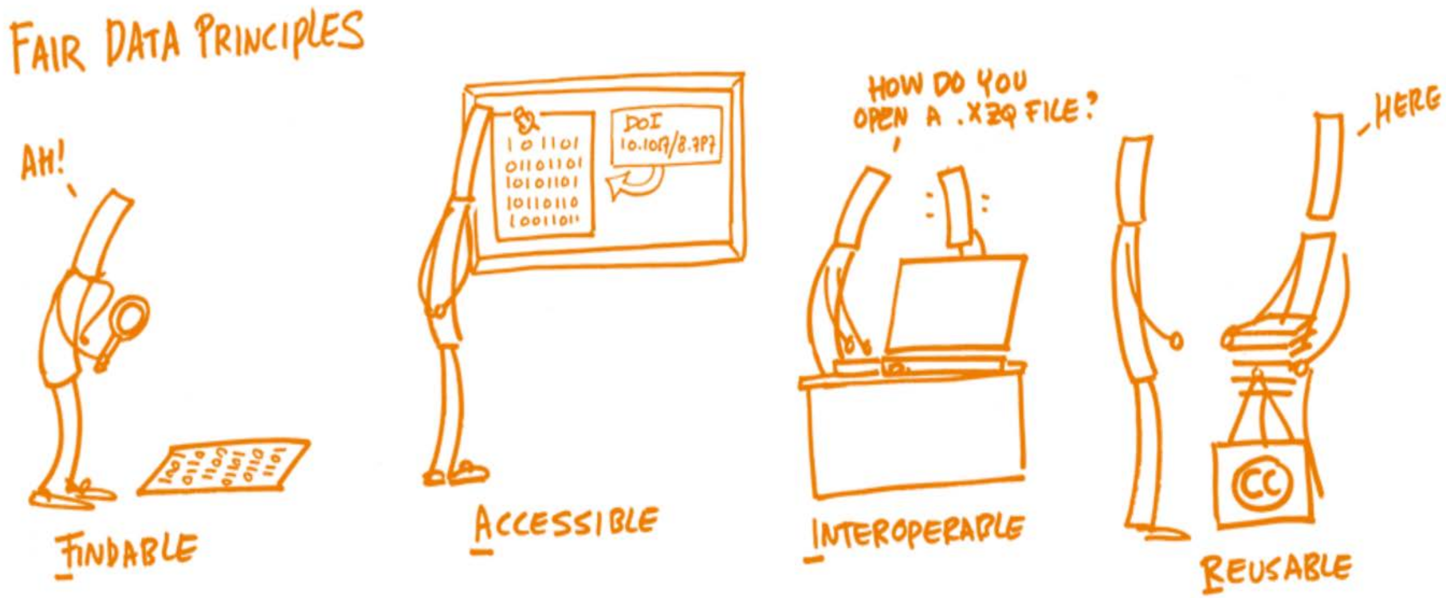


Image: https://book.fosteropenscience.eu

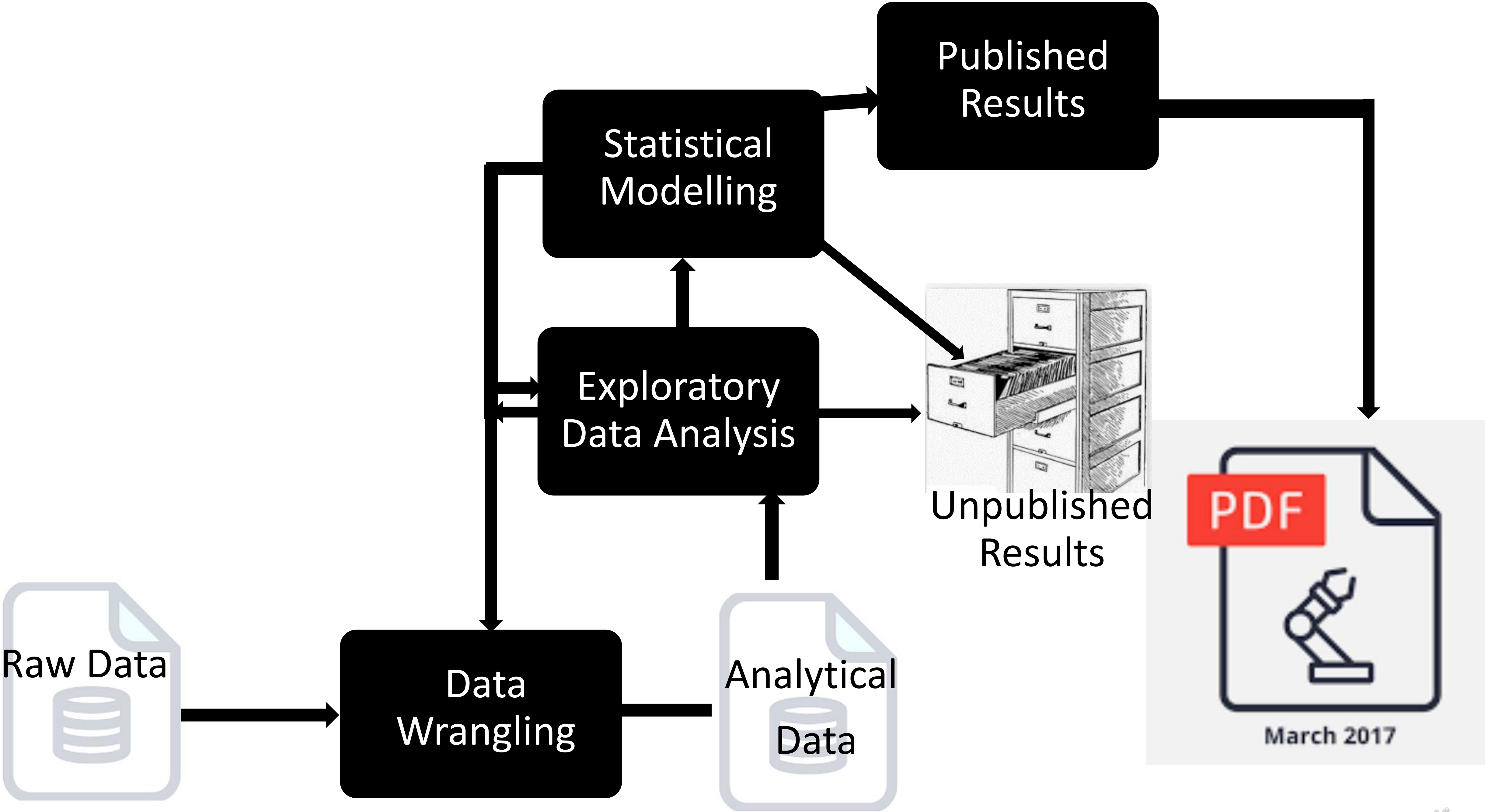
GitHub

OSFHOME

Conclusions

© DYN3D/ISTOCKPHOTO.COM

Royalty-free stock photo ID: 223869256



Published Results

Statistical Modelling

Exploratory Data Analysis

Unpublished Results

Raw Data

Data Wrangling

Analytical Data

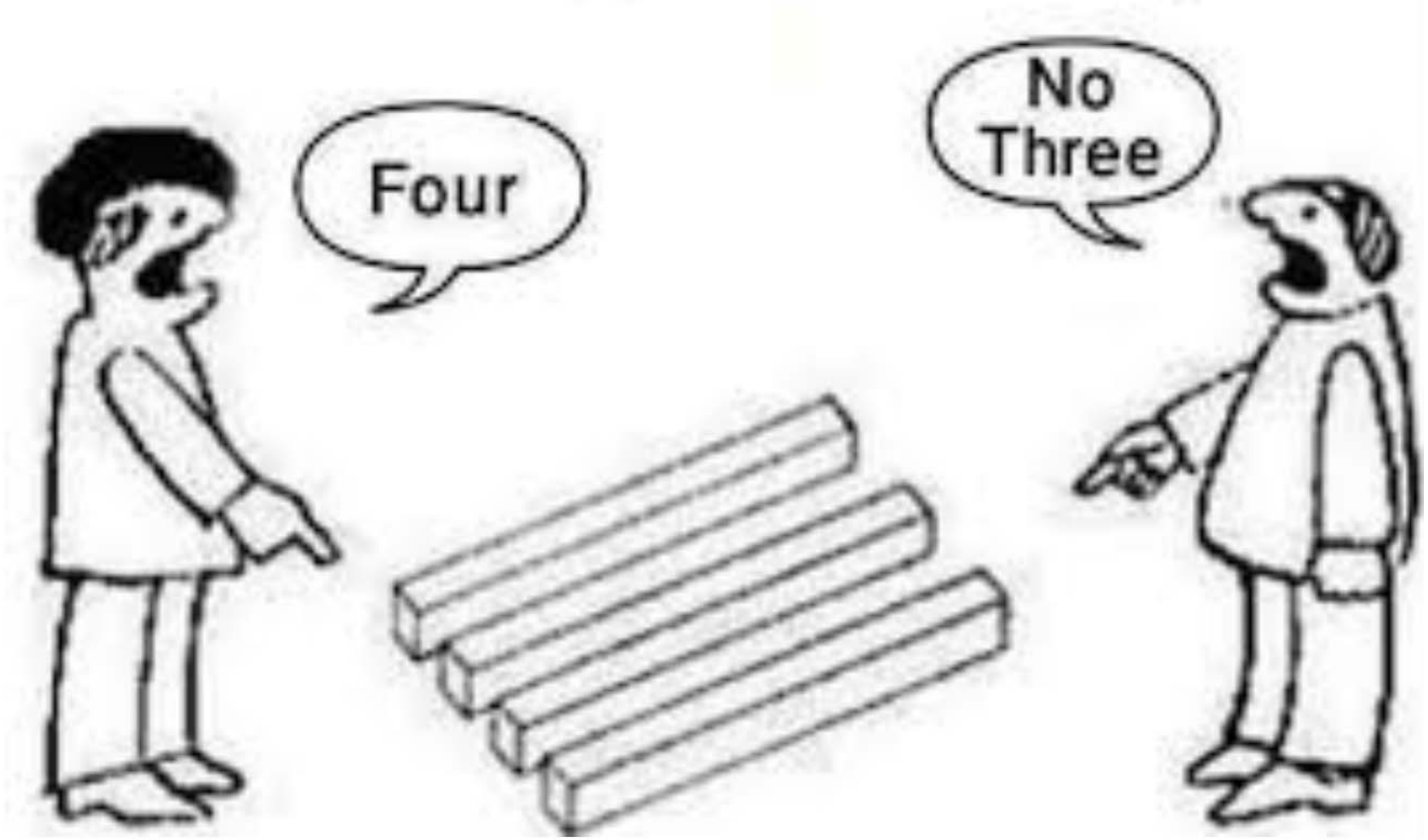
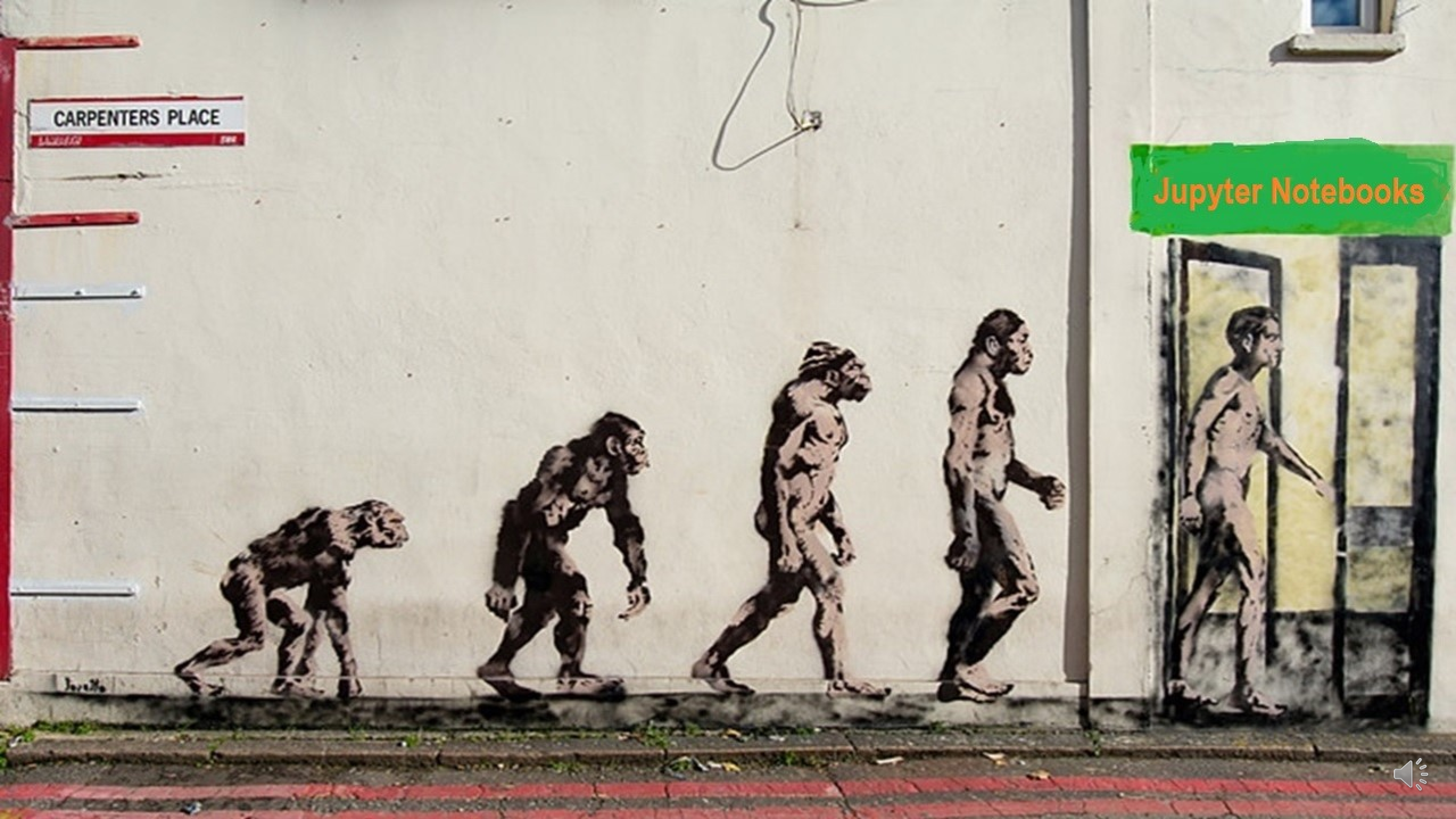


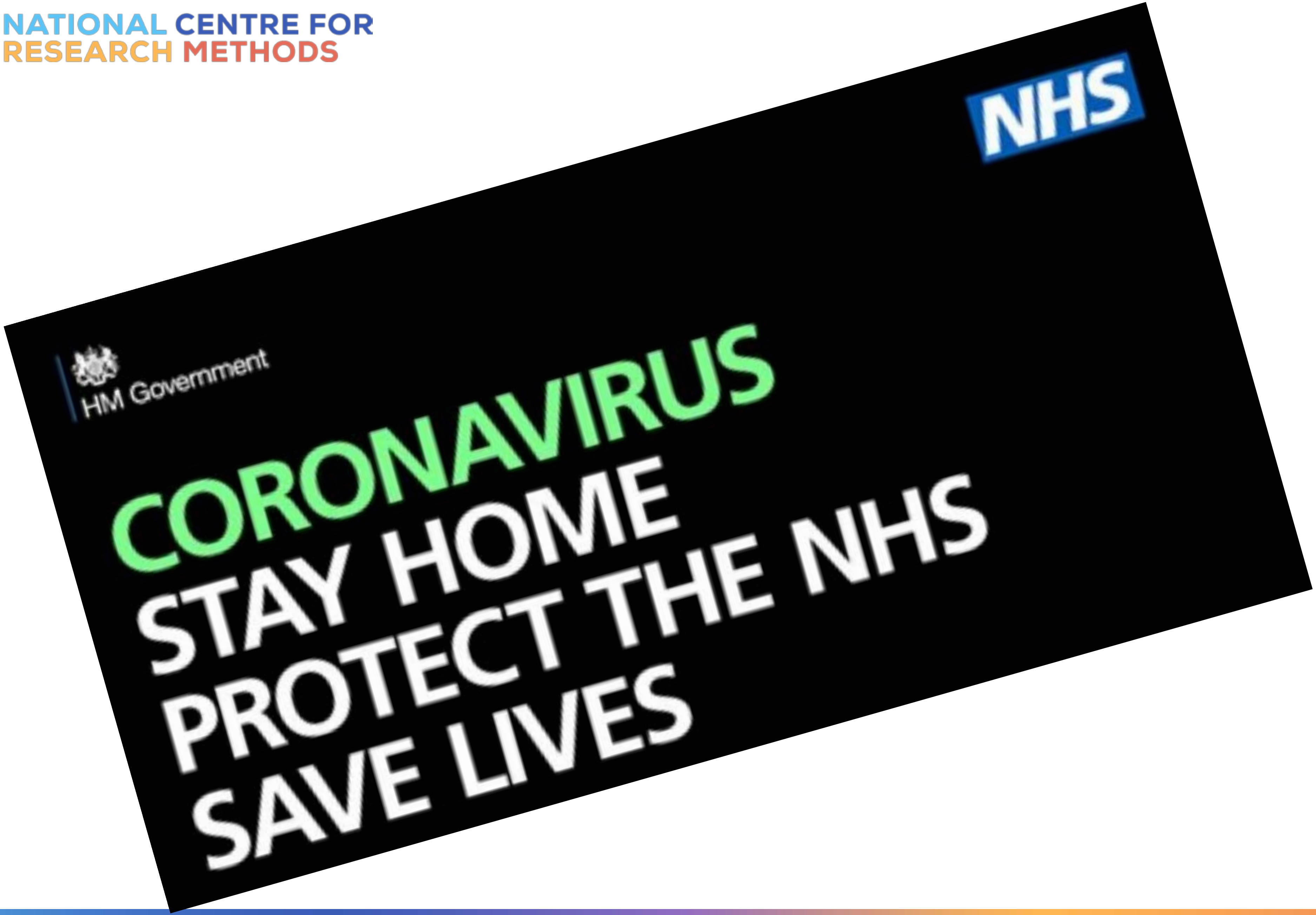
Image https://qph.fs.quoracdn.net/main-qimg-88ae2063a1267ea7de76968a488dd723

Image: https://i0.wp.com/www.bayesianspectacles.org/wp-content/uploads/2018/01/TexasSharpShooter.png?zoom=2&fit=357%2C300&ssl=1

PREREGISTERED



Jupyter Notebooks



NHS

HM Government

CORONAVIRUS
STAY HOME
PROTECT THE NHS
SAVE LIVES

How to cite this video

Gayle, V. (2020) *Reproducible Data Analysis.* Available at: https://www.ncrm.ac.uk (Accessed: day month year)

NATIONAL CENTRE FOR RESEARCH METHODS

Reproducible Data Analysis

Professor Vernon Gayle
vernon.gayle@ed.ac.uk
@Profbigvern
https://github.com/vernongayle

UKRI Economic and Social Research Council
UNIVERSITY OF Southampton
MANCHESTER 1824 The University of Manchester
THE UNIVERSITY OF EDINBURGH