Visualizing Logistic Regression:

Application of coloring book technique in a reproducible *ggplot2* system

Andriy Koval, Ph.D.

Department of Statistics & Data Science

Colloquium Series

2018-11-08-Friday

Orlando, Florida

UCF UNIVERSITY OF CENTRAL FLORIDA

github.com/andkov/ipdln-2018-hackathon

Today

- 1. Introduce a graphing technique “*coloring book*”
- 2. Demonstrate a production workflow for its implementation
- 3. Build a case for reproducible projects

About myself

- Ph.D. in Quantitative Methods, Psychology (2014)
- Reproducible research enthusiast since 2012
- Graph maker
- See work at https://github.com/andkov
- These slides and more at http://andriy.rbind.io

MIDDLE TENNESSEE STATE UNIVERSITY

The UNIVERSITY of OKLAHOMA

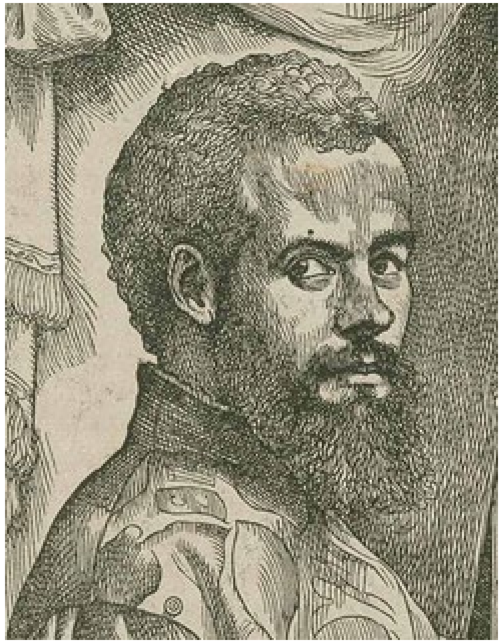
VANDERBILT

University of Victoria

UBC THE UNIVERSITY OF BRITISH COLUMBIA

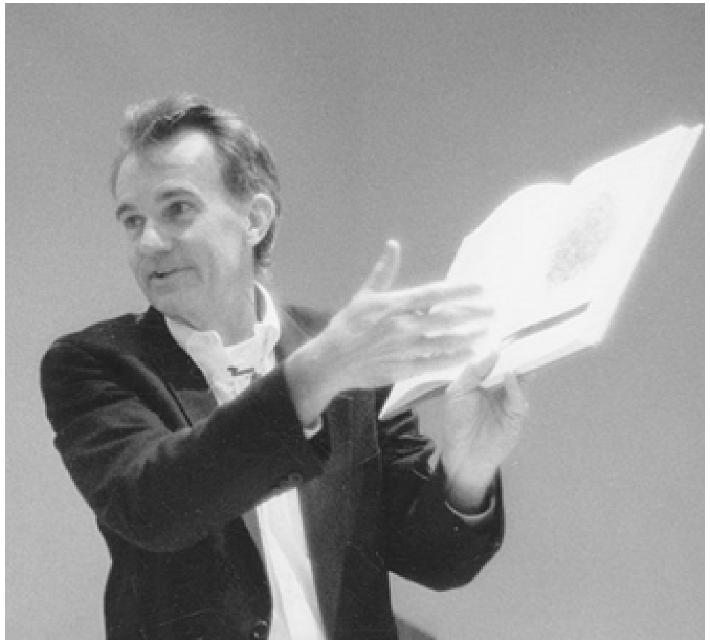
UCF UNIVERSITY OF CENTRAL FLORIDA

Key influences



Andreas Vesalius

John Tukey



Edward Tufte

Hadley Wickham

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon

0.1 Modeling form: univariate logistic regression with categorical predictors

0.2 Graphical form: faceted scatterplot in ggplot2

0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans

1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

INTERNATIONAL Population Data Linkage NETWORK

https://www.ipdln.org/

IJPDS International Journal of Population Data Science

The Science of Data About People

IBM

Statistics Canada Statistique Canada

September 11, 2018

- *Event* : Linked Data Innovation Challenge
- *Data* : Synthetic mortality data
- *Records* : 4,346,649
- *Variables* : 34

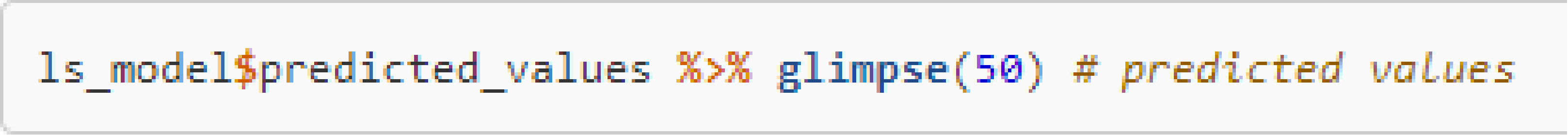
Banff, Alberta

Q: What explains mortality among immigrants?

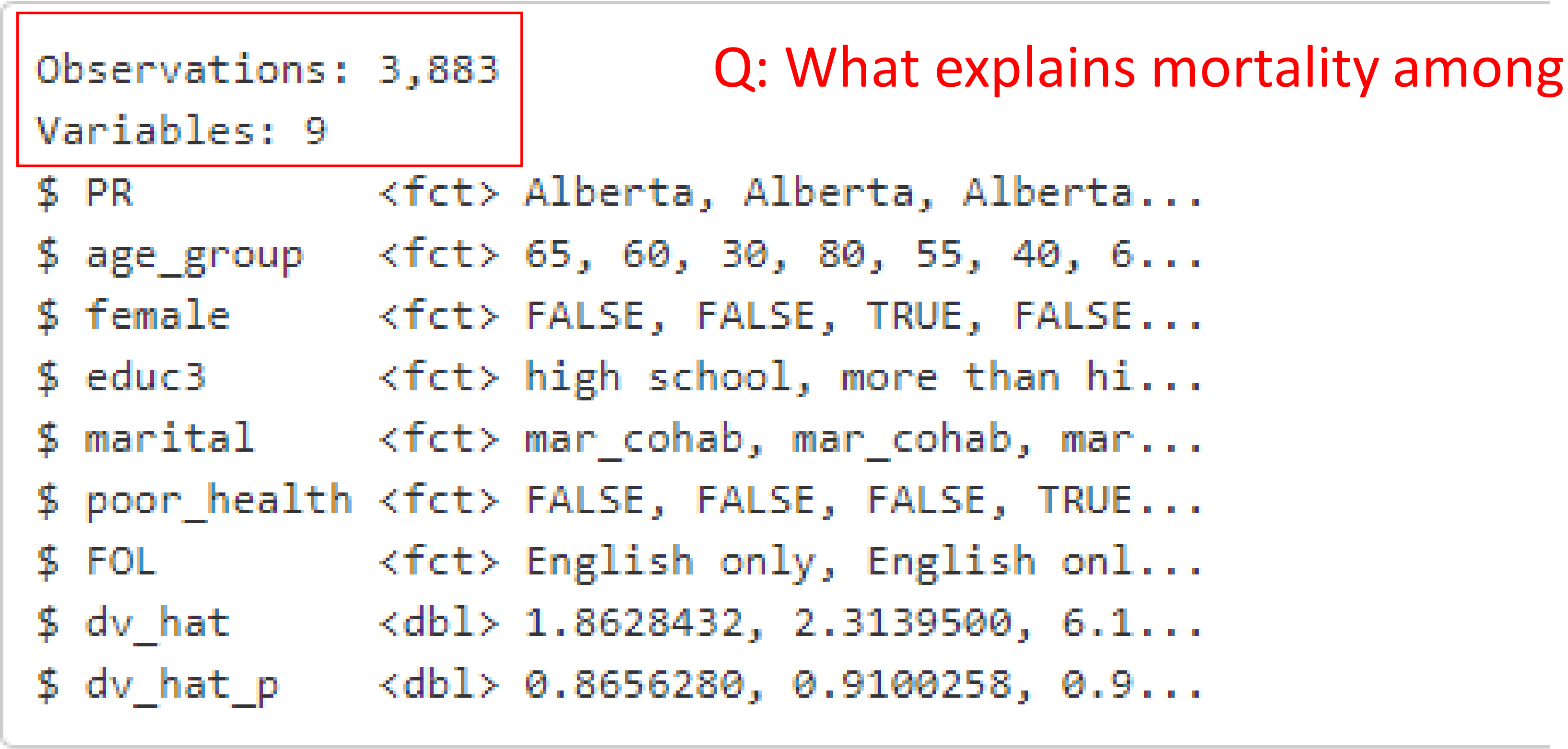
github.com/andkov/ipdln-2018-hackathon

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff



```
ls_model$predicted_values %>% glimpse(50) # predicted values
```



```
Observations: 3,883
Variables: 9
$ PR          <fct> Alberta, Alberta, Alberta...
$ age_group   <fct> 65, 60, 30, 80, 55, 40, 6...
$ female      <fct> FALSE, FALSE, TRUE, FALSE...
$ educ3       <fct> high school, more than hi...
$ marital     <fct> mar_cohab, mar_cohab, mar...
$ poor_health <fct> FALSE, FALSE, FALSE, TRUE...
$ FOL         <fct> English only, English onl...
$ dv_hat      <dbl> 1.8628432, 2.3139500, 6.1...
$ dv_hat_p    <dbl> 0.8656280, 0.9100258, 0.9...
```

Q: What explains mortality among immigrants?

Originally:

Number of records: 4,346,649
Number of variables: 34

Data recreated from model parameters based on a stratified sample (N=1000) from 4 provinces

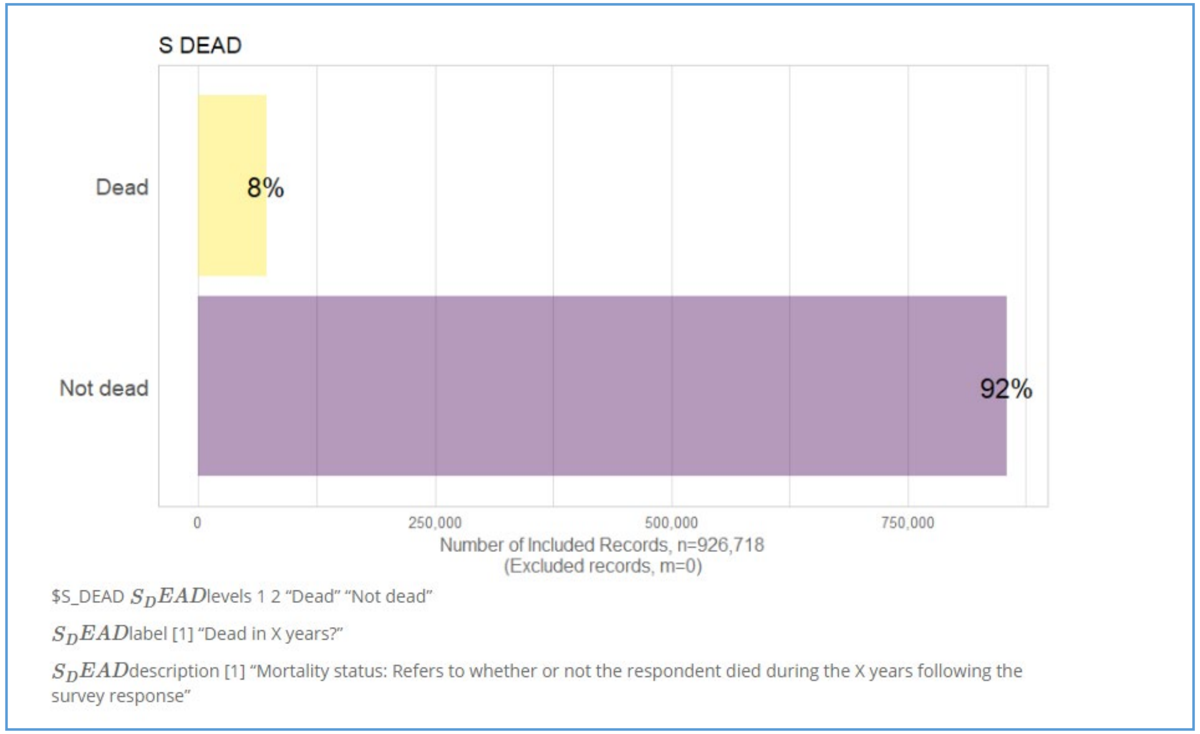
You can use this data to recreate the graphs from this talk
with the script *./reports/graphing-phase-only/graphing-phase-only.R*
Clone github.com/andkov/ipdln-2018-hackathon for better experience

A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Dead in X years



$S_DEAD $S_{D}EAD$levels 1 2 “Dead” “Not dead”

$S_{D}EAD$label [1] “Dead in X years?”

$S_{D}EAD$description [1] “Mortality status: Refers to whether or not the respondent died during the X years following the survey response”

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

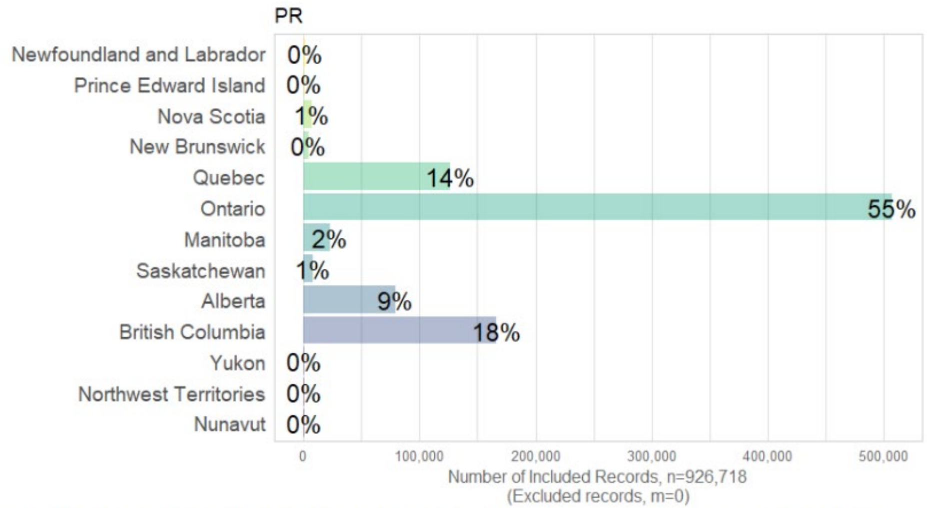
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Province of residence



$PR PRlevels 10 11 12 “Newfoundland and Labrador” “Prince Edward Island” “Nova Scotia” 13 24 35 “New Brunswick” “Quebec” “Ontario” 46 47 48 “Manitoba” “Saskatchewan” “Alberta” 59 60 61 “British Columbia” “Yukon” “Northwest Territories” 62 “Nunavut”

PRlabel [1] “Province of residence”

PRdescription [1] “Province or territory of residence”

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

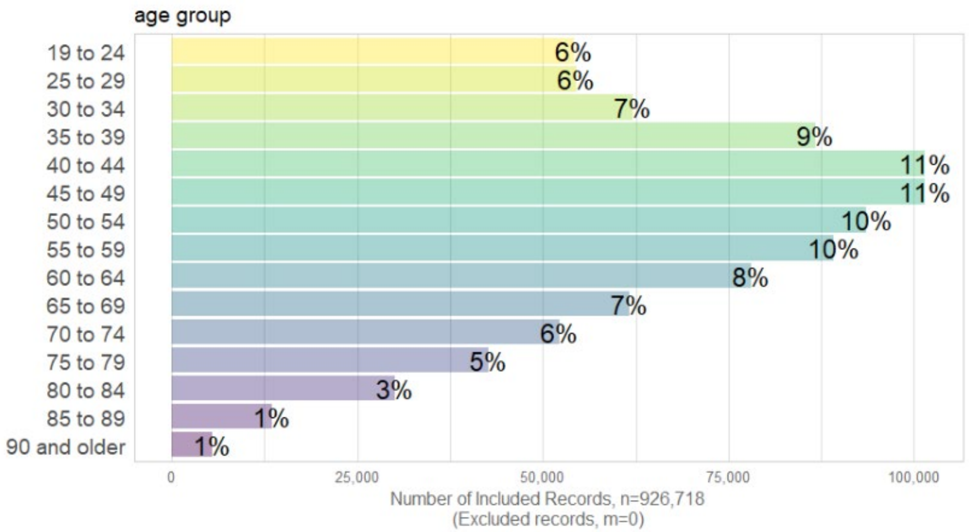
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

5-year age category



\$age_group $age_{g}roup$levels 1 2 3 4 5 6 “19 to 24” “25 to 29” “30 to 34” “35 to 39” “40 to 44” “45 to 49” 7 8 9 10 11 12 “50 to 54” “55 to 59” “60 to 64” “65 to 69” “70 to 74” “75 to 79” 13 14 15 “80 to 84” “85 to 89” “90 and older”

$age_{g}roup$label [1] “Age”

$age_{g}roup$description [1] “Age: grouped”

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

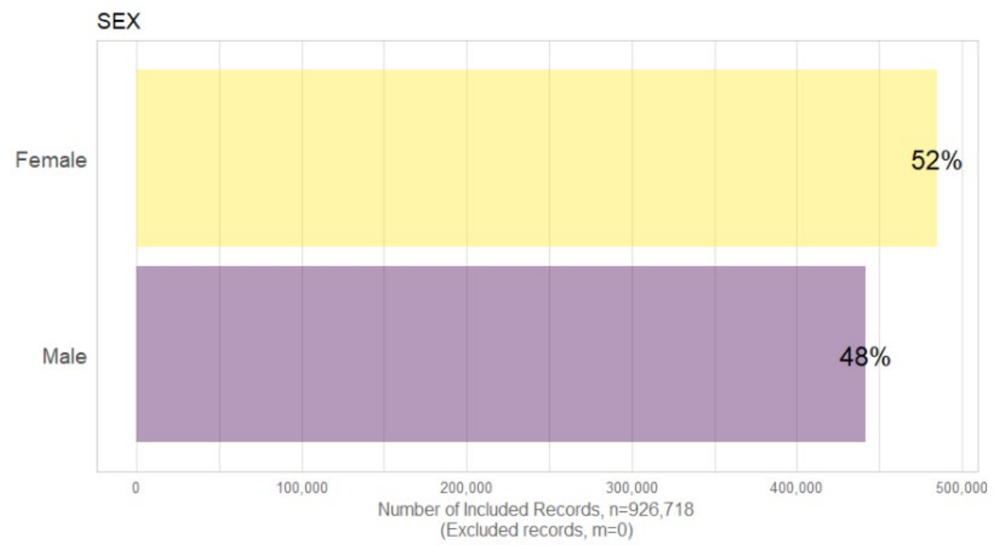
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Sex



$SEX *SEX*levels 1 2 “Female” “Male”

*SEX*label [1] “Sex”

*SEX*description [1] “Sex”

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

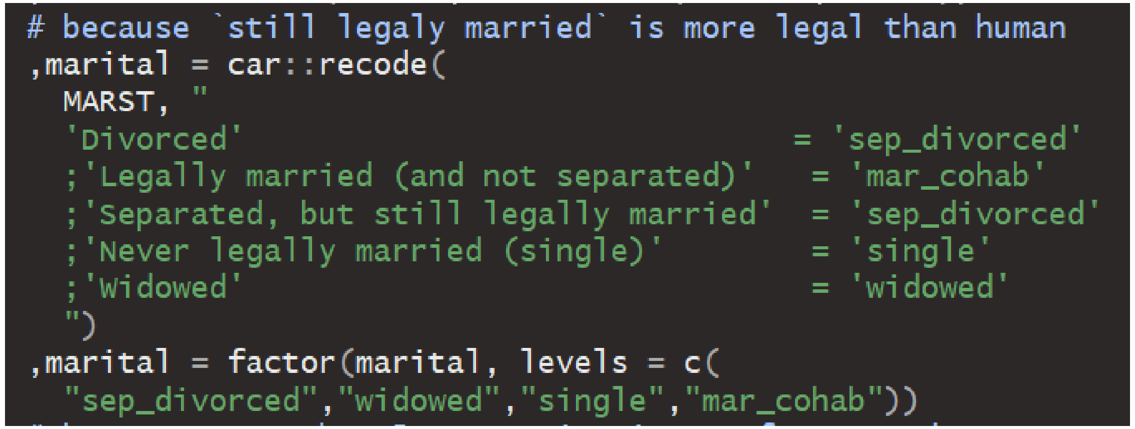
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

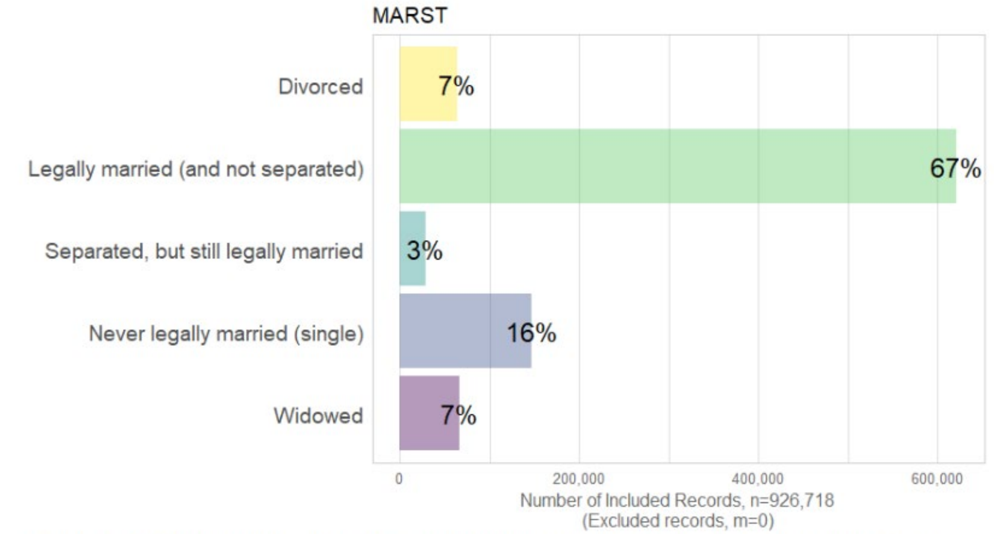
0.1 Modeling form

$$dv \sim -1 + PR + age\_group + female + \boxed{marital} + educ3 + poor\_health + FOL$$

Marital Status



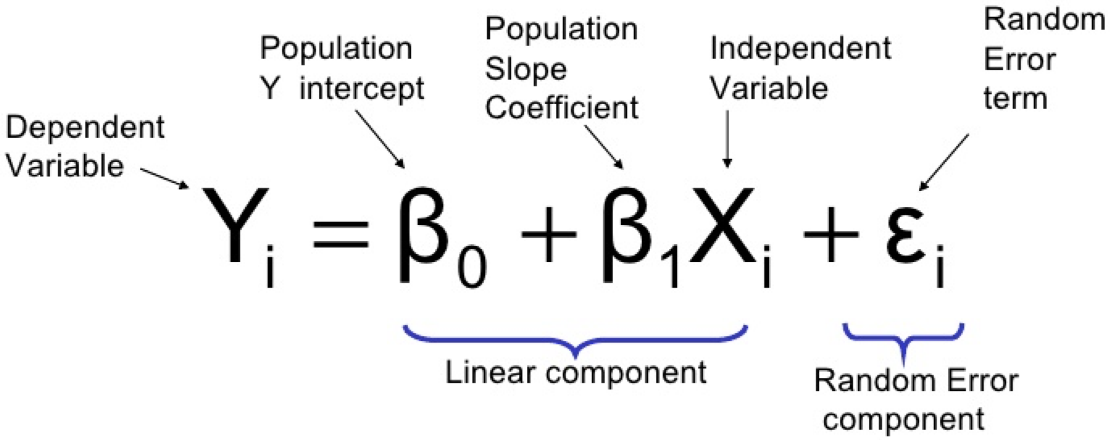
```
# because `still legaly married` is more legal than human
,marital = car::recode(
  MARST, "
  'Divorced'                               = 'sep_divorced'
  ;'Legally married (and not separated)'   = 'mar_cohab'
  ;'Separated, but still legally married'  = 'sep_divorced'
  ;'Never legally married (single)'        = 'single'
  ;'Widowed'                               = 'widowed'
  ")
,marital = factor(marital, levels = c(
  "sep_divorced","widowed","single","mar_cohab"))
```



$MARST $MARST$levels 1 2 “Divorced” “Legally married (and not separated)” 3 4 “Separated, but still legally married” “Never legally married (single)” 5 “Widowed”

$MARST$label [1] “Marital status”

$MARST$description [1] “Marital Status: Refers to the legal marital status of the person.”



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

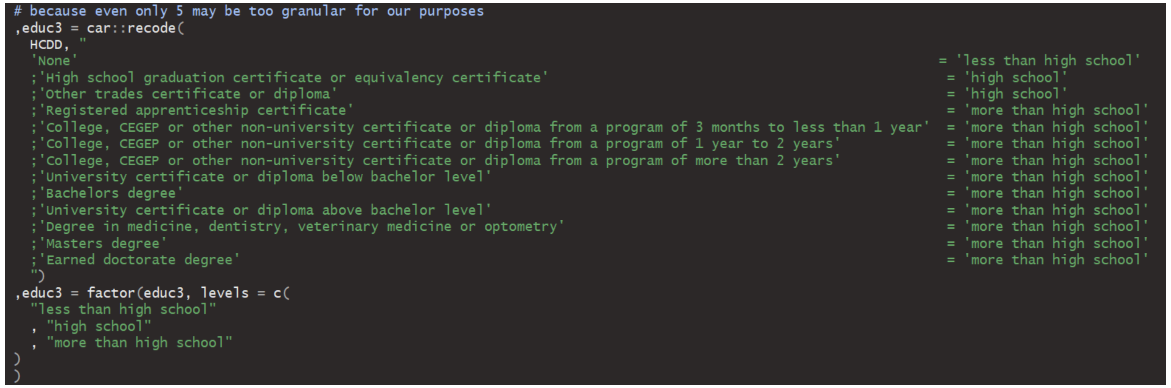
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

A. Graphing Technique

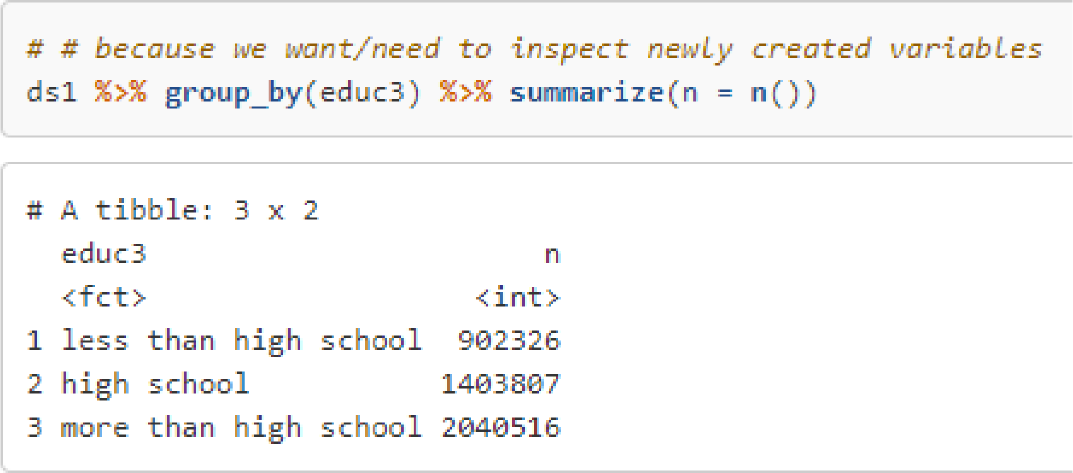
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Highest Degree

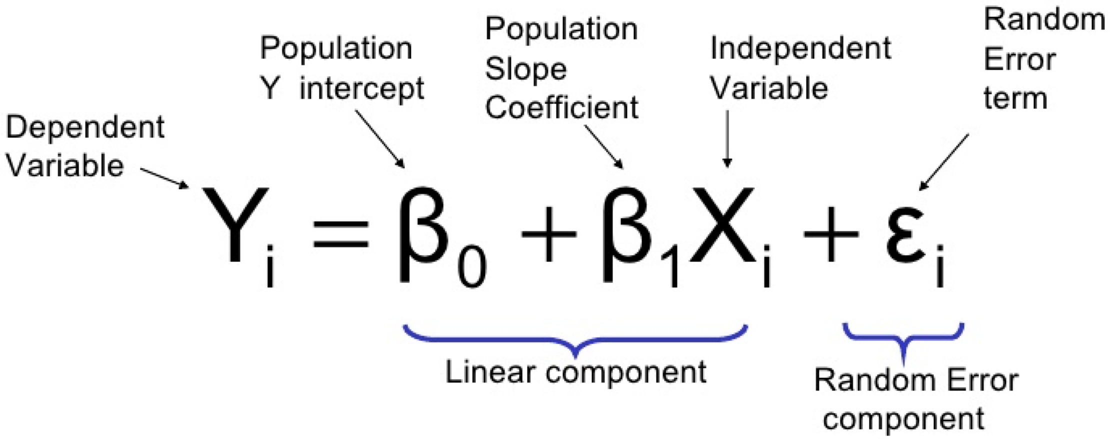


```
# because even only 5 may be too granular for our purposes
,educ3 = car::recode(
  HCDD, "
  'None'                                                                                                          = 'less than high school'
  ;'High school graduation certificate or equivalency certificate'                                                 = 'high school'
  ;'Other trades certificate or diploma'                                                                           = 'high school'
  ;'Registered apprenticeship certificate'                                                                         = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year'  = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years'             = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of more than 2 years'             = 'more than high school'
  ;'University certificate or diploma below bachelor level'                                                        = 'more than high school'
  ;'Bachelors degree'                                                                                              = 'more than high school'
  ;'University certificate or diploma above bachelor level'                                                        = 'more than high school'
  ;'Degree in medicine, dentistry, veterinary medicine or optometry'                                                = 'more than high school'
  ;'Masters degree'                                                                                                = 'more than high school'
  ;'Earned doctorate degree'                                                                                       = 'more than high school'
  ")
,educ3 = factor(educ3, levels = c(
  "less than high school"
  , "high school"
  , "more than high school"
)
)
```



```
# # because we want/need to inspect newly created variables
ds1 %>% group_by(educ3) %>% summarize(n = n())
```

```
# A tibble: 3 x 2
  educ3                       n
  <fct>                   <int>
1 less than high school  902326
2 high school           1403807
3 more than high school 2040516
```



$$\underbrace{Y_i = \beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

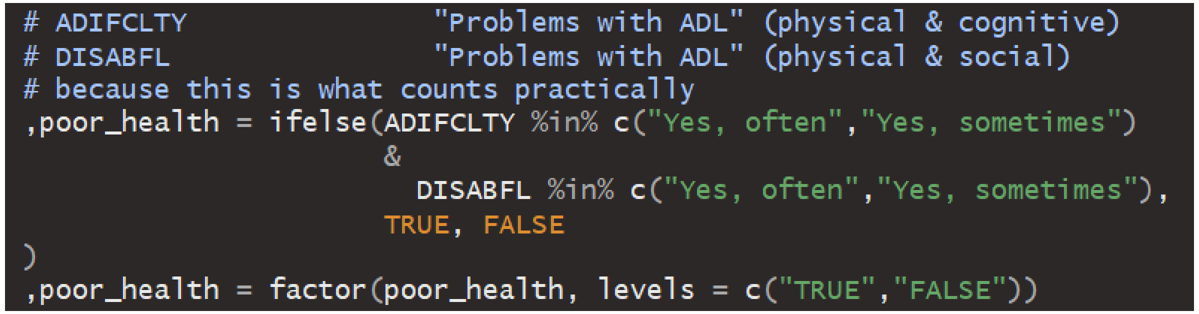
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

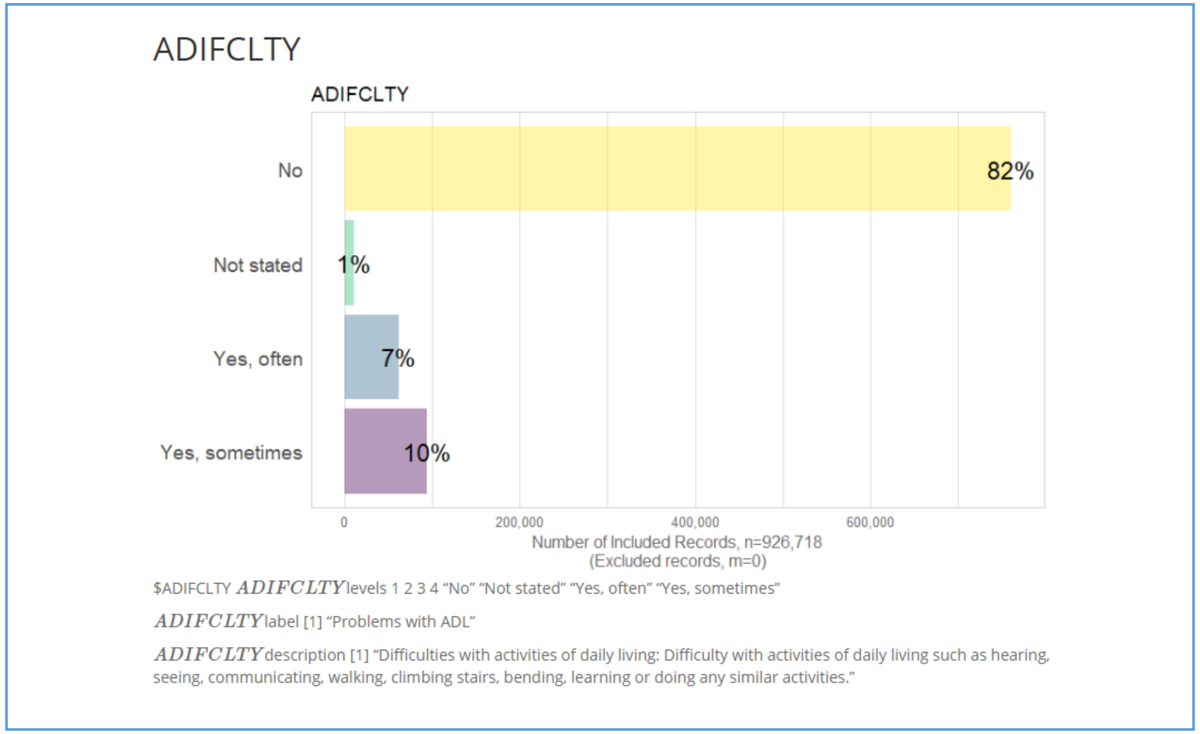
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living



```
# ADIFCLTY                "Problems with ADL" (physical & cognitive)
# DISABFL                 "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```

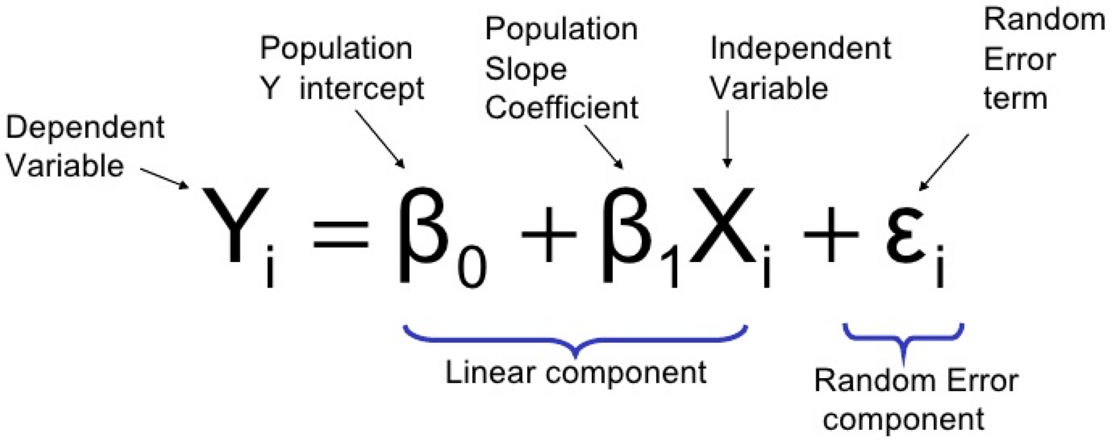


ADIFCLTY

$ADIFCLTY *ADIFCLTY* levels 1 2 3 4 “No” “Not stated” “Yes, often” “Yes, sometimes”

ADIFCLTY label [1] “Problems with ADL”

ADIFCLTY description [1] “Difficulties with activities of daily living: Difficulty with activities of daily living such as hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing any similar activities.”



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$\underbrace{Y_i = \beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

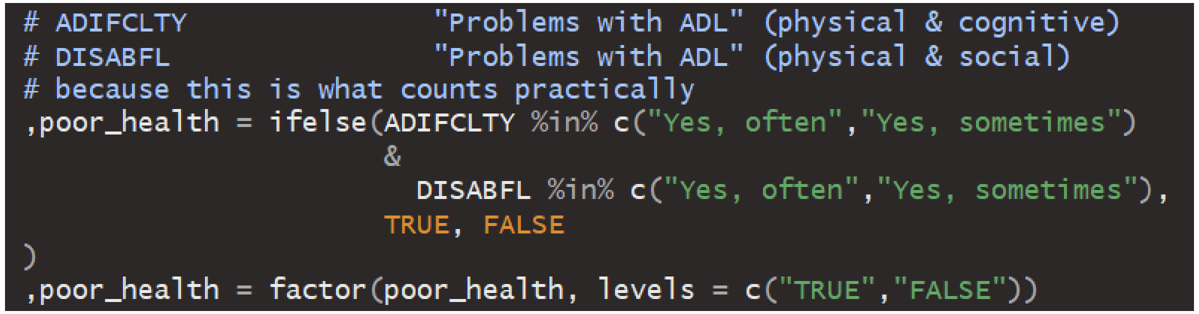
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

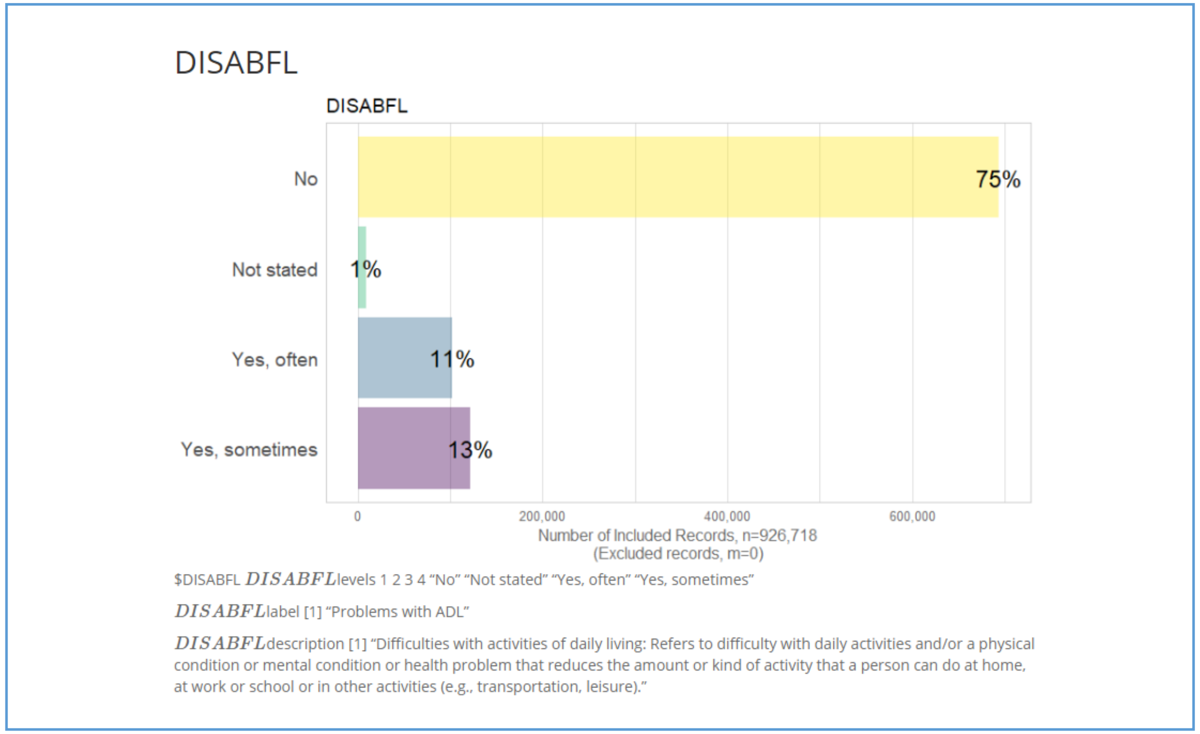
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living



```
# ADIFCLTY               "Problems with ADL" (physical & cognitive)
# DISABFL                "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```



$DISABFL *DISABFL*levels 1 2 3 4 “No” “Not stated” “Yes, often” “Yes, sometimes”

*DISABFL*label [1] “Problems with ADL”

*DISABFL*description [1] “Difficulties with activities of daily living: Refers to difficulty with daily activities and/or a physical condition or mental condition or health problem that reduces the amount or kind of activity that a person can do at home, at work or school or in other activities (e.g., transportation, leisure).”

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

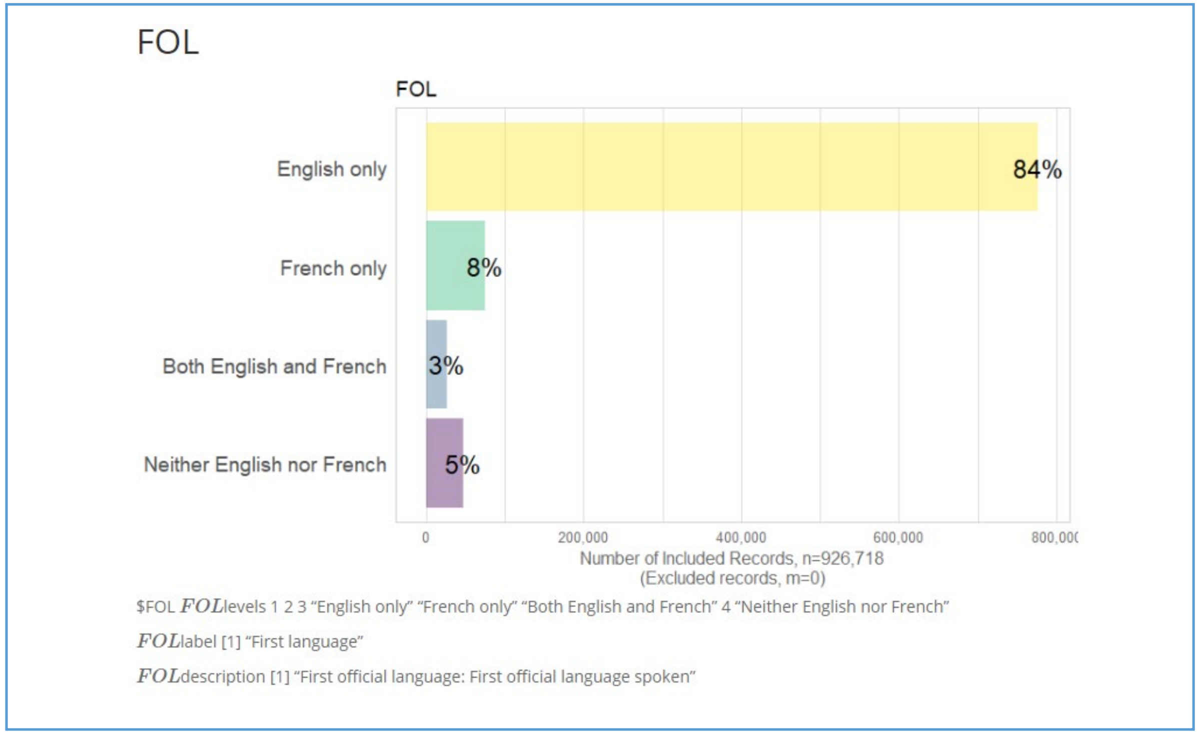
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

First Official Language



$FOL *FOL*levels 1 2 3 “English only” “French only” “Both English and French” 4 “Neither English nor French”

*FOL*label [1] “First language”

*FOL*description [1] “First official language: First official language spoken”

Dependent Variable

Population Y intercept

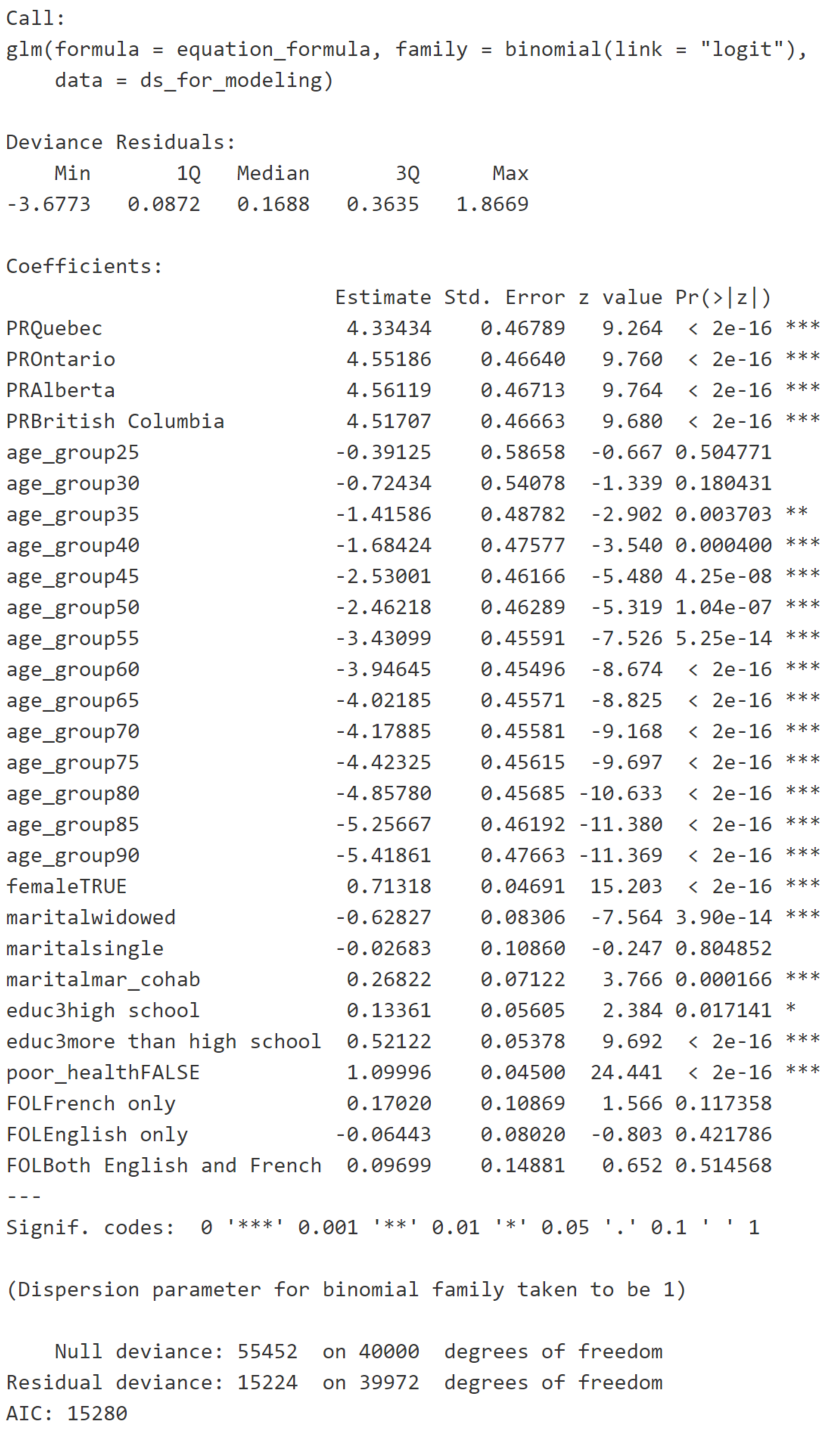
Population Slope Coefficient

Independent Variable

Random Error term

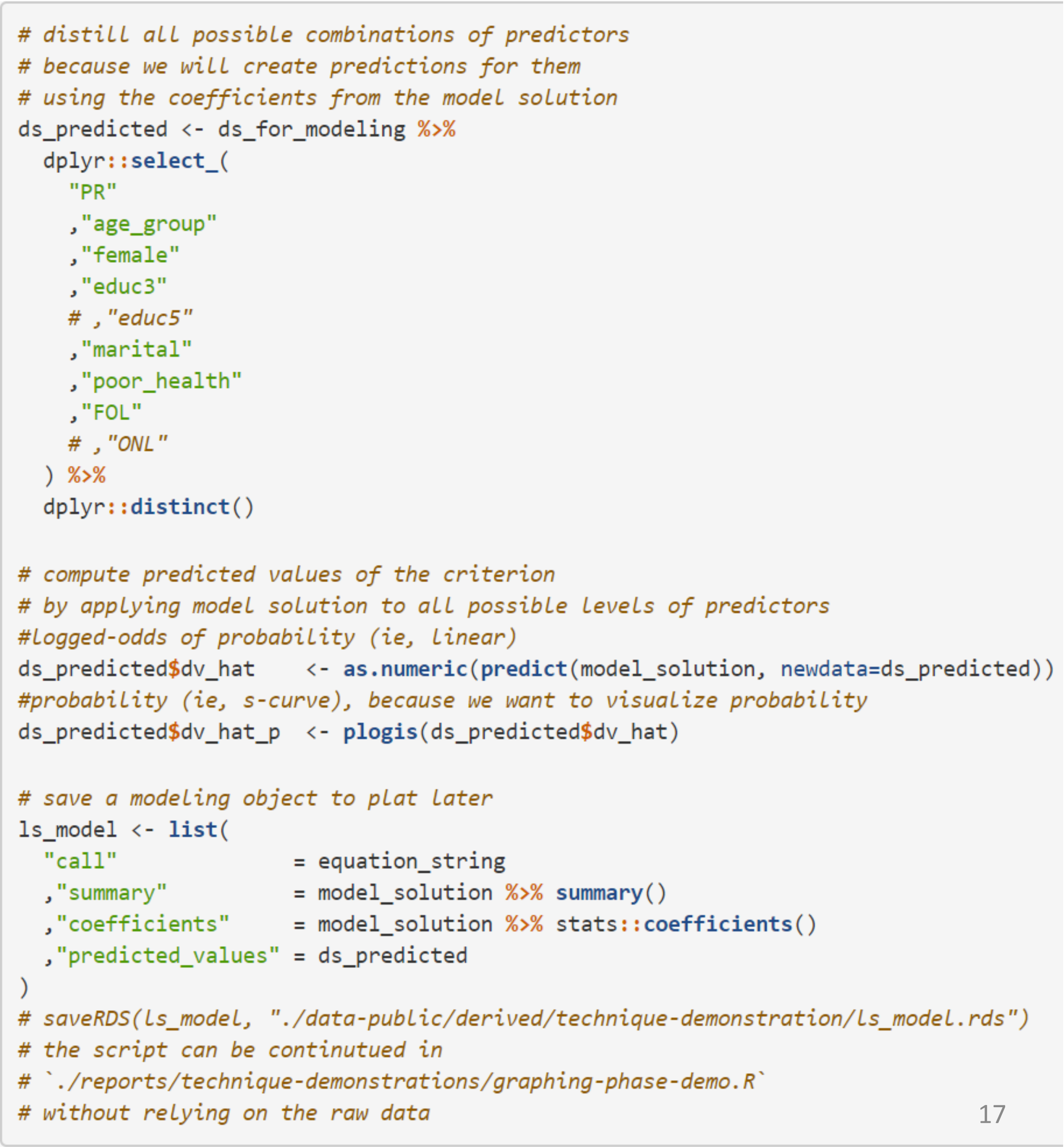
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$



```
Call:
glm(formula = equation_formula, family = binomial(link = "logit"), 
    data = ds_for_modeling)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.6773   0.0872   0.1688   0.3635   1.8669  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
PRQuebec                    4.33434    0.46789   9.264  < 2e-16 ***
PROntario                   4.55186    0.46640   9.760  < 2e-16 ***
PRAlberta                   4.56119    0.46713   9.764  < 2e-16 ***
PRBritish Columbia          4.51707    0.46663   9.680  < 2e-16 ***
age_group25                -0.39125    0.58658  -0.667 0.504771    
age_group30                -0.72434    0.54078  -1.339 0.180431    
age_group35                -1.41586    0.48782  -2.902 0.003703 ** 
age_group40                -1.68424    0.47577  -3.540 0.000400 ***
age_group45                -2.53001    0.46166  -5.480 4.25e-08 ***
age_group50                -2.46218    0.46289  -5.319 1.04e-07 ***
age_group55                -3.43099    0.45591  -7.526 5.25e-14 ***
age_group60                -3.94645    0.45496  -8.674  < 2e-16 ***
age_group65                -4.02185    0.45571  -8.825  < 2e-16 ***
age_group70                -4.17885    0.45581  -9.168  < 2e-16 ***
age_group75                -4.42325    0.45615  -9.697  < 2e-16 ***
age_group80                -4.85780    0.45685 -10.633  < 2e-16 ***
age_group85                -5.25667    0.46192 -11.380  < 2e-16 ***
age_group90                -5.41861    0.47663 -11.369  < 2e-16 ***
femaleTRUE                  0.71318    0.04691  15.203  < 2e-16 ***
maritalwidowed             -0.62827    0.08306  -7.564 3.90e-14 ***
maritalsingle              -0.02683    0.10860  -0.247 0.804852    
maritalmar_cohab            0.26822    0.07122   3.766 0.000166 ***
educ3high school            0.13361    0.05605   2.384 0.017141 *  
educ3more than high school  0.52122    0.05378   9.692  < 2e-16 ***
poor_healthFALSE            1.09996    0.04500  24.441  < 2e-16 ***
FOLFrench only              0.17020    0.10869   1.566 0.117358    
FOLEnglish only            -0.06443    0.08020  -0.803 0.421786    
FOLBoth English and French  0.09699    0.14881   0.652 0.514568    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55452  on 40000  degrees of freedom
Residual deviance: 15224  on 39972  degrees of freedom
AIC: 15280
```

Model Prediction



```r
# distill all possible combinations of predictors
# because we will create predictions for them
# using the coefficients from the model solution
ds_predicted <- ds_for_modeling %>% 
  dplyr::select_(
    "PR"
    ,"age_group"
    ,"female"
    ,"educ3"
    # ,"educ5"
    ,"marital"
    ,"poor_health"
    ,"FOL"
    # ,"ONL"
  ) %>% 
  dplyr::distinct()

# compute predicted values of the criterion
# by applying model solution to all possible levels of predictors
#logged-odds of probability (ie, linear)
ds_predicted$dv_hat    <- as.numeric(predict(model_solution, newdata=ds_predicted))
#probability (ie, s-curve), because we want to visualize probability
ds_predicted$dv_hat_p  <- plogis(ds_predicted$dv_hat)

# save a modeling object to plat later
ls_model <- list(
  "call"               = equation_string
  ,"summary"           = model_solution %>% summary()
  ,"coefficients"      = model_solution %>% stats::coefficients()
  ,"predicted_values" = ds_predicted
)
# saveRDS(ls_model, "./data-public/derived/technique-demonstration/ls_model.rds")
# the script can be continutued in
# `./reports/technique-demonstrations/graphing-phase-demo.R`
# without relying on the raw data
```

A. Graphing Technique

0.2 Graphical form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

LEGEND

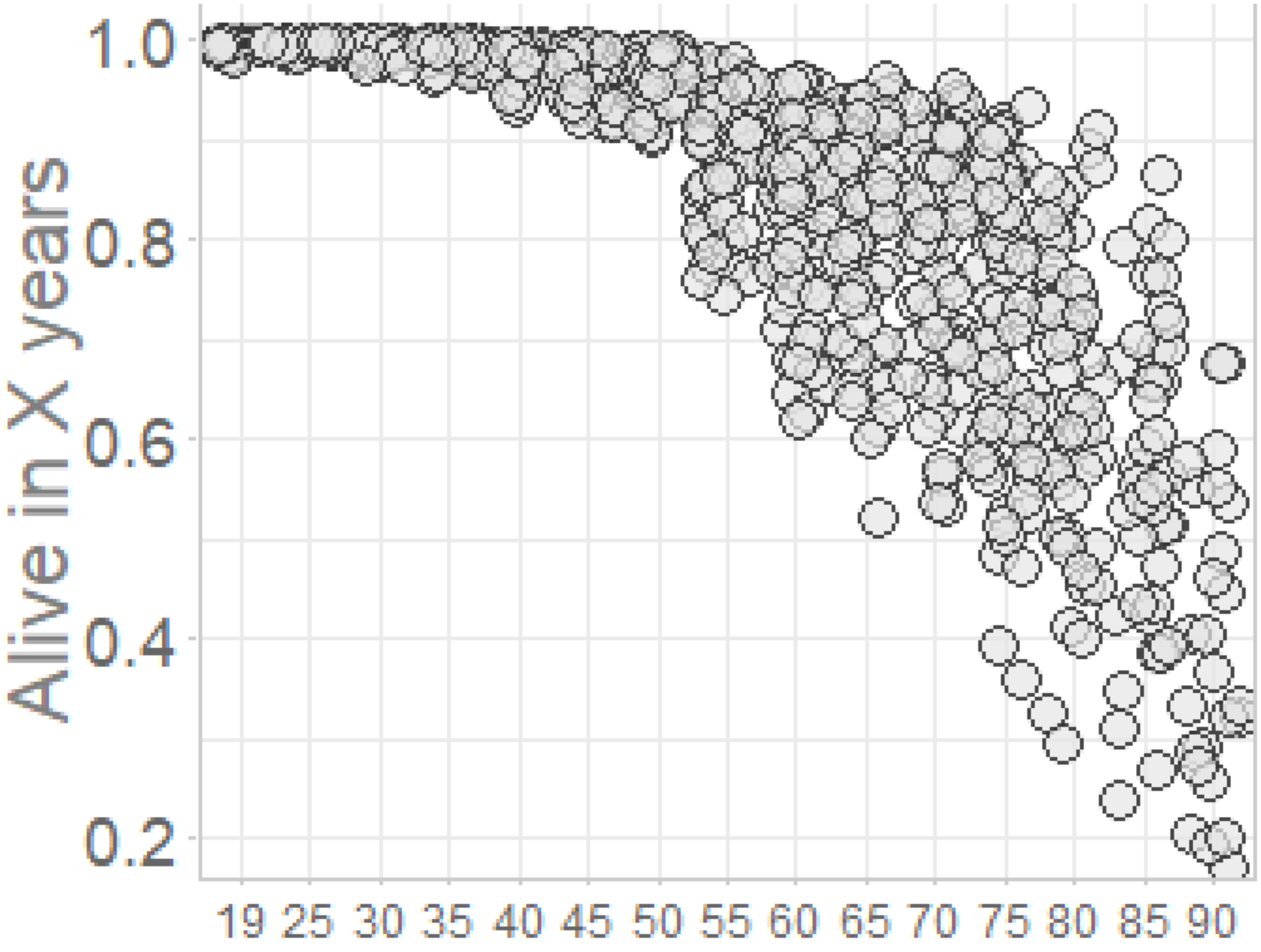
point = person

Y-axis = probability R is dead in X years

X-axis = age group (floor of 5-year category)

The higher the dot = the higher the chance to be alive in X years

Visualizing probability instead of log-odds because it is more intuitive



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

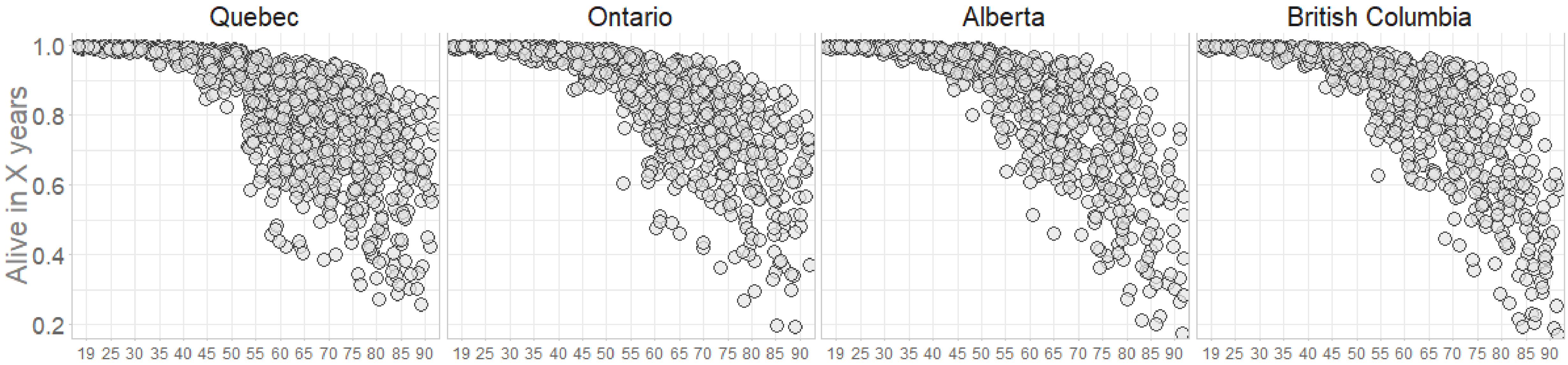
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

A. Graphing Technique

0.2 Graphical form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



LEGEND

Facet = Province of residence

A. Graphing Technique

0.2 Graphical form

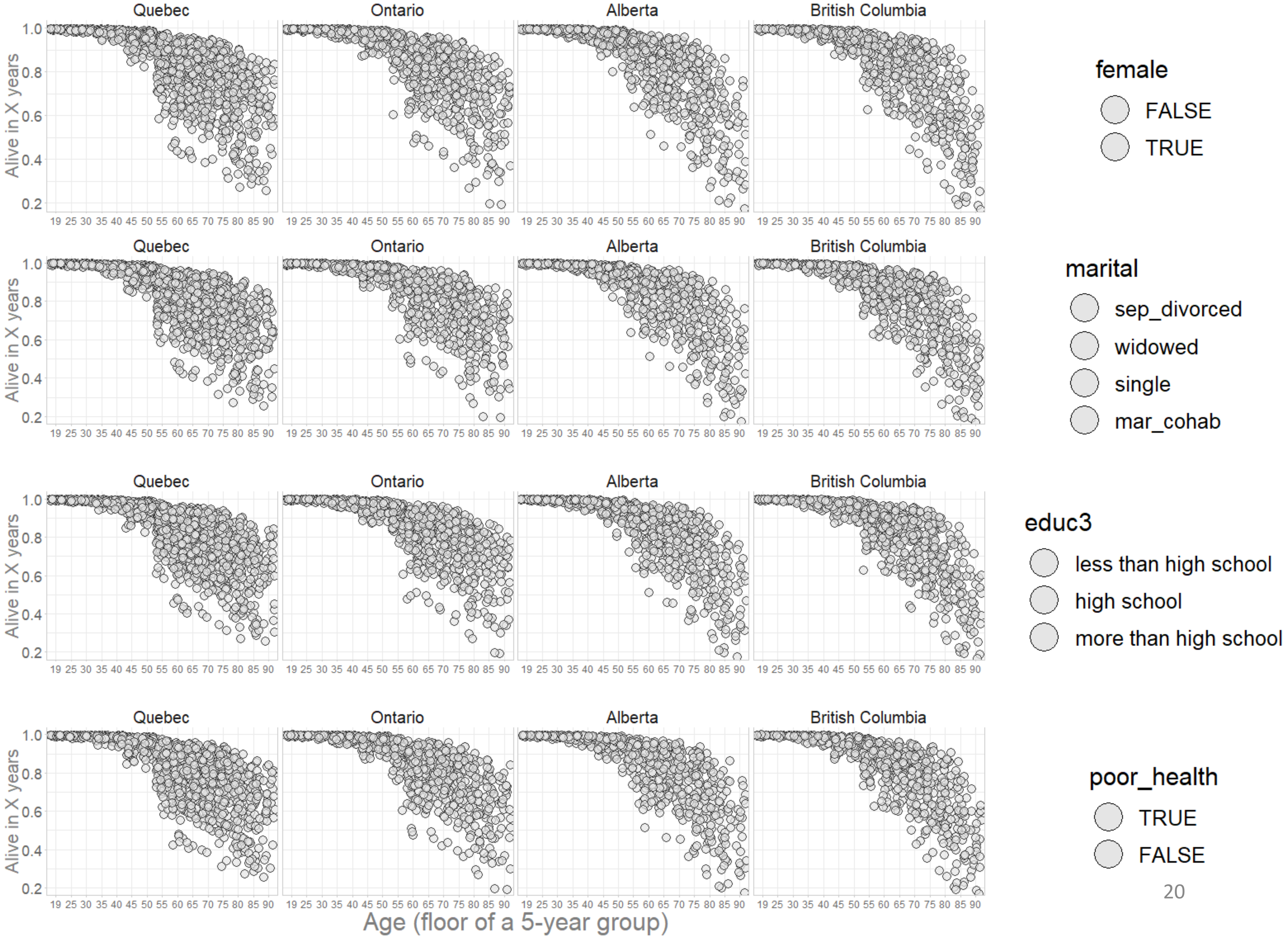
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

LEGEND

Rows = duplicate of each other (for now).

Notice that FOL is not displayed

The book is ready for coloring



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

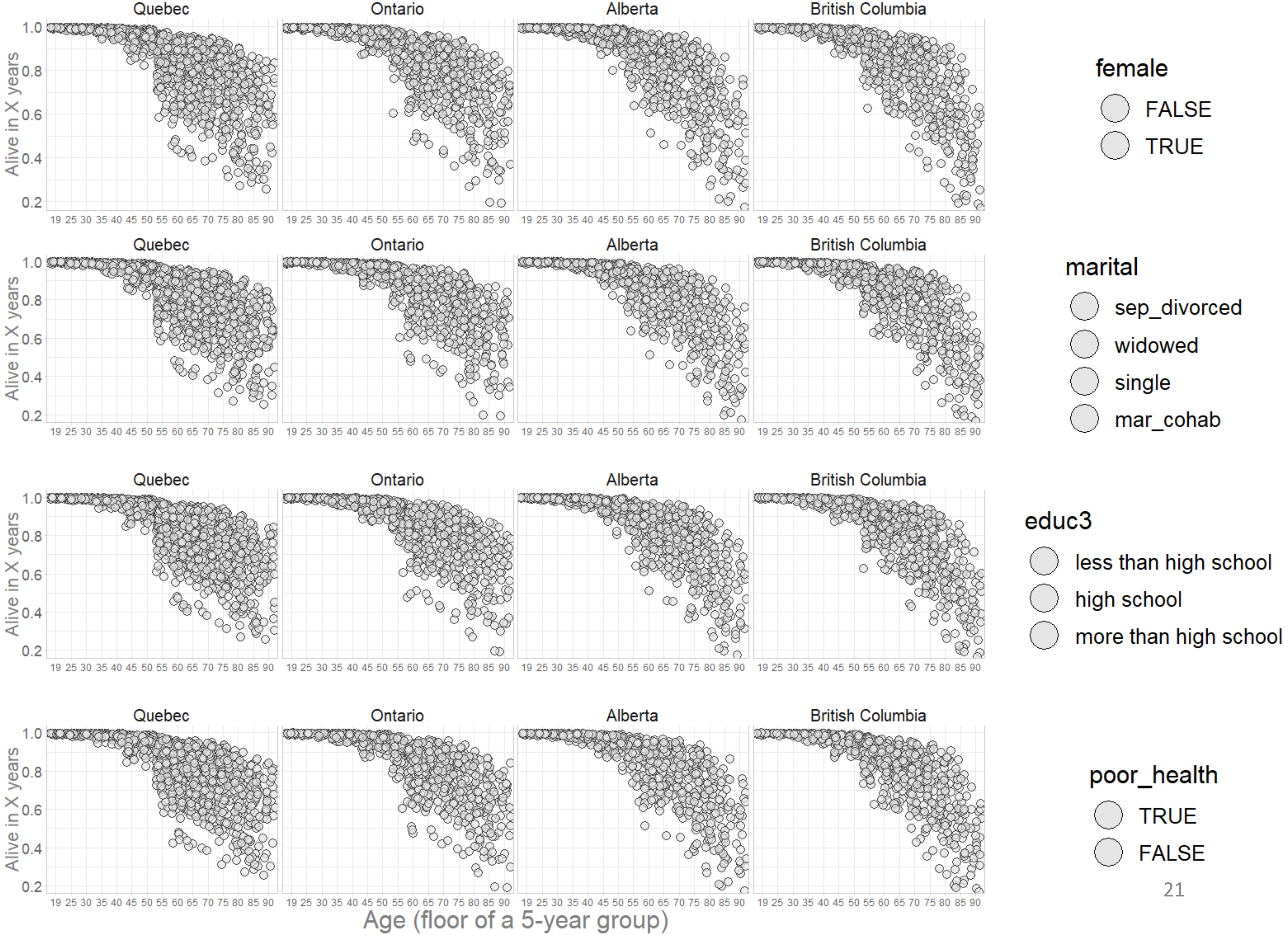
What should the “reference group” be for each predictor?

What do we expect based on existing research?

Informed expectation

Reference group

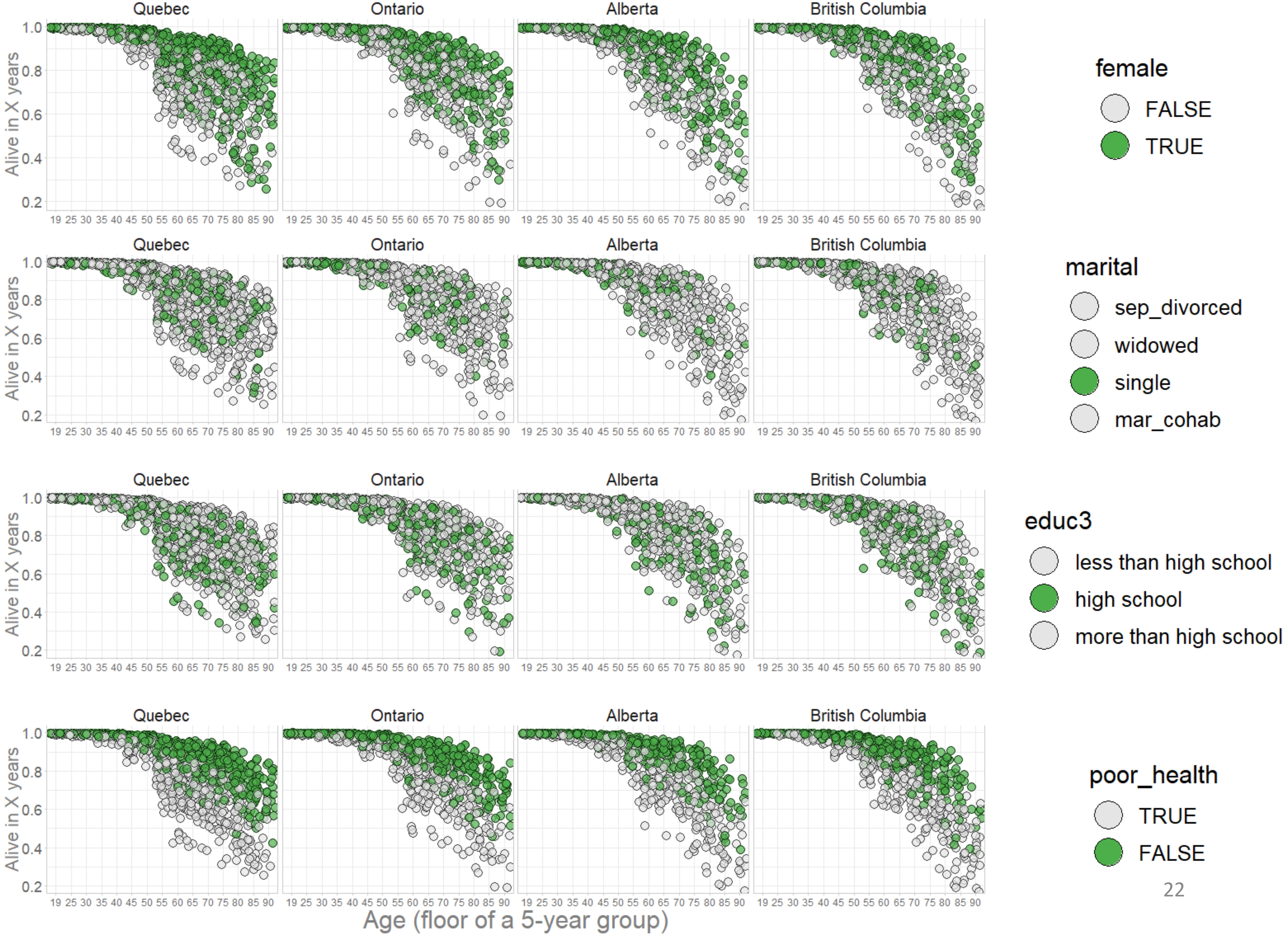
?



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



Informed expectation

Reference group

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

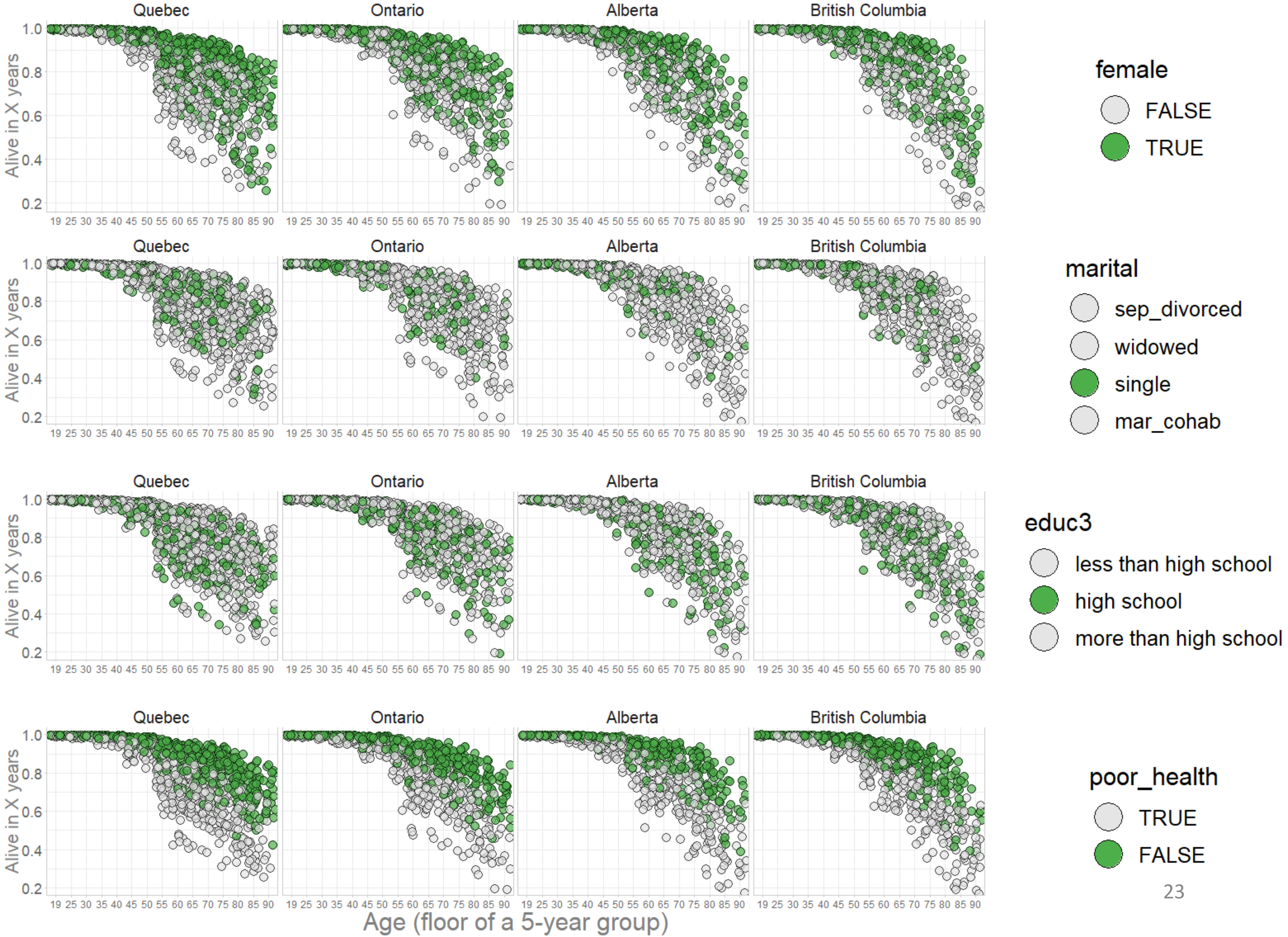
QUESTION

Compared to reference group, what levels of predictors are expected to **increase** the mortality risk?

Informed expectation

Moderately increased risk ?

Reference group



A. Graphing Technique

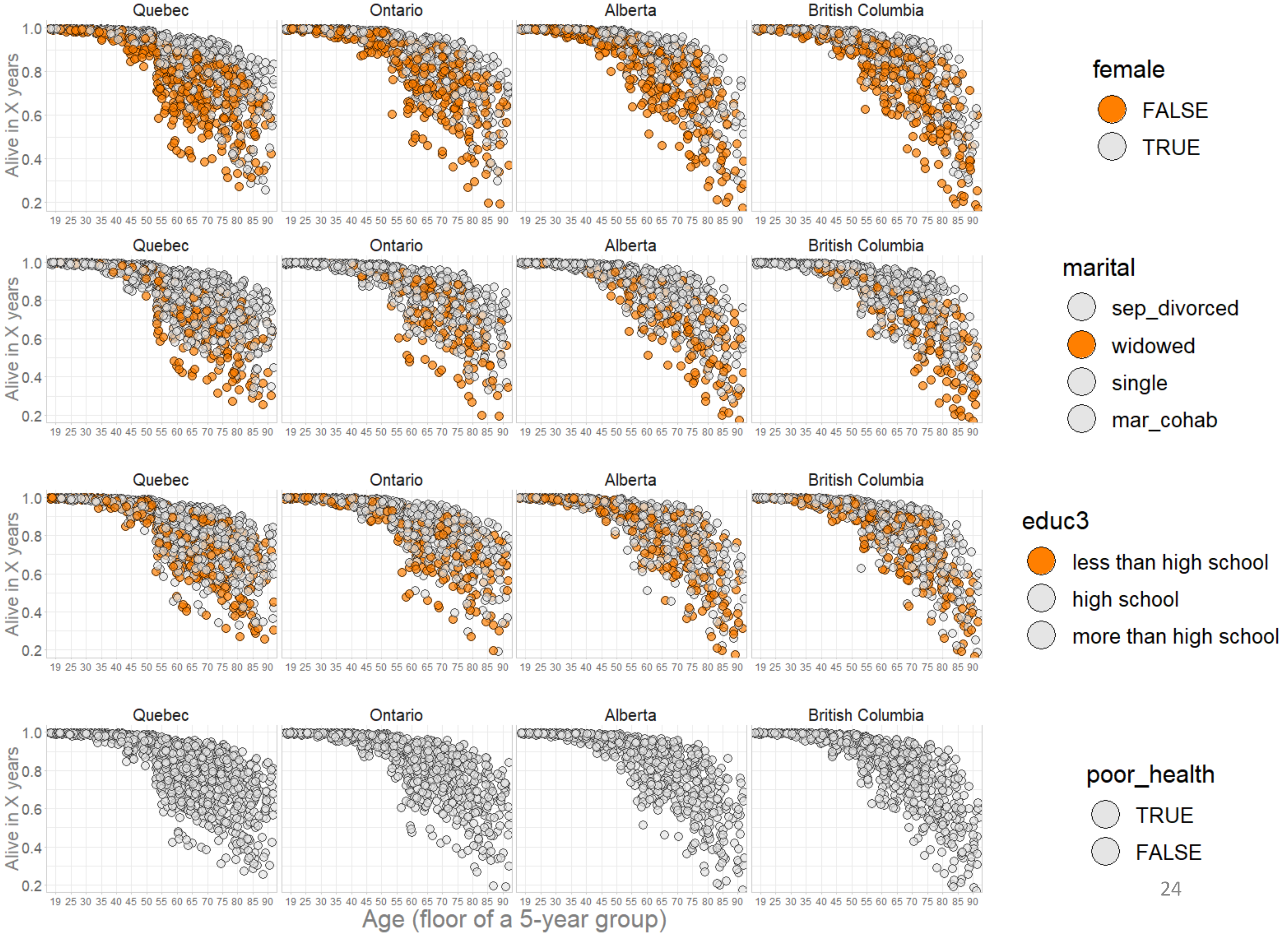
0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Informed expectation

Moderately increased risk

Reference group



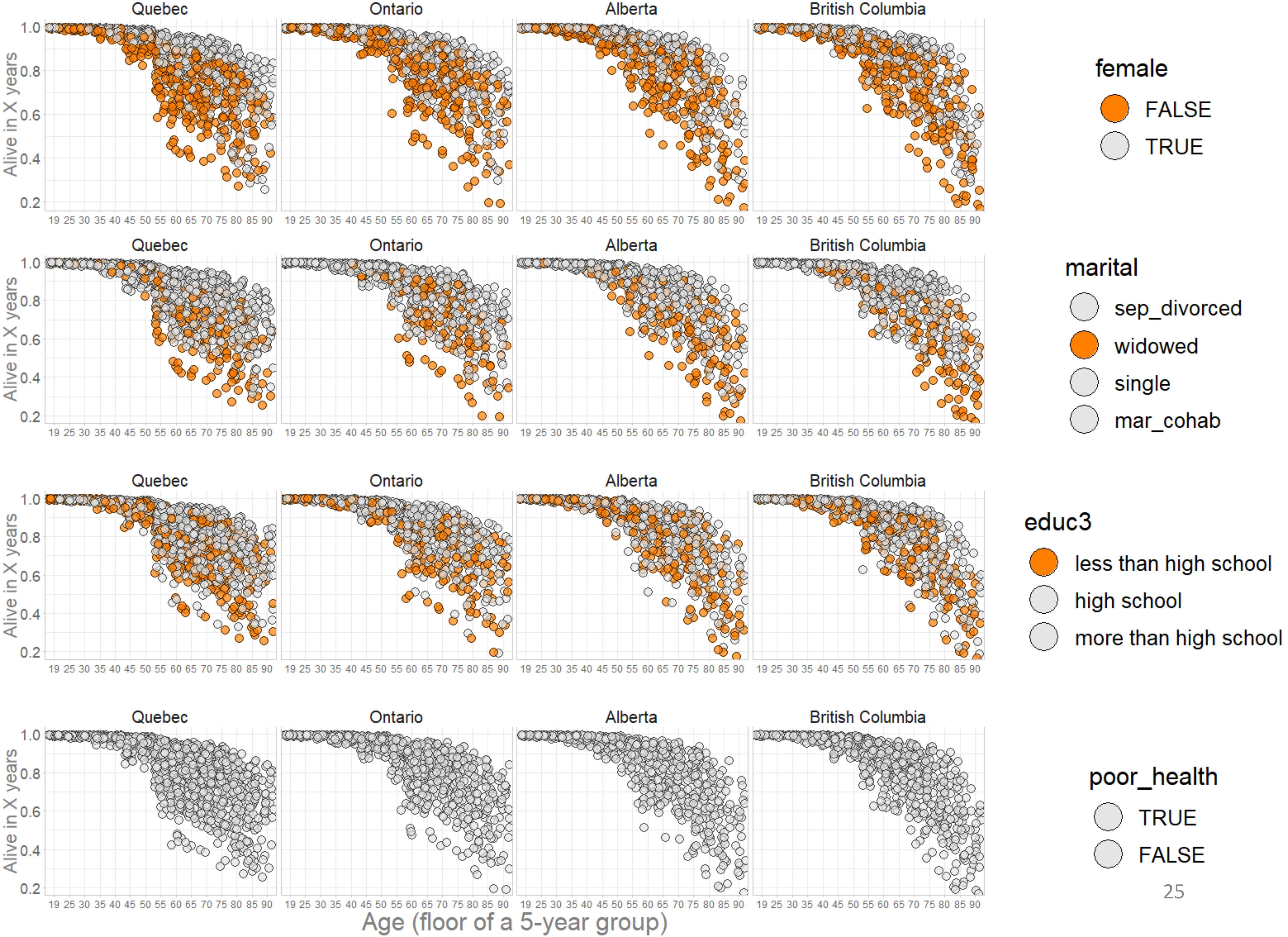
A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

Compared to reference group, what levels of predictors are expected to **decrease** the mortality risk?



Informed expectation

Moderately increased risk

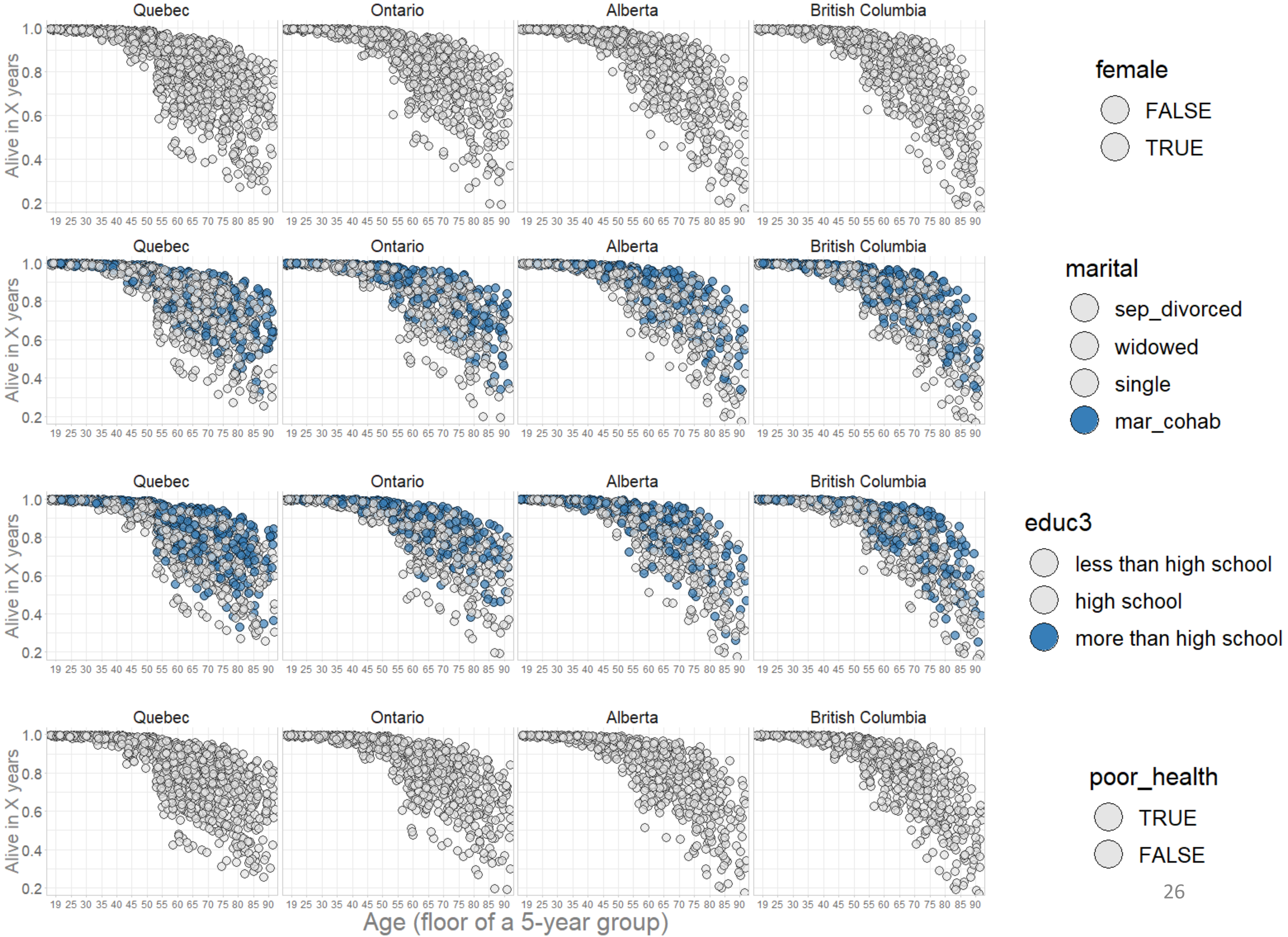
Reference group

Moderately decreased risk ?

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



Informed expectation

Moderately increased risk

Reference group

Moderately decreased risk

A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

What levels of predictors are expected to affect mortality risk drastically?

Informed expectation

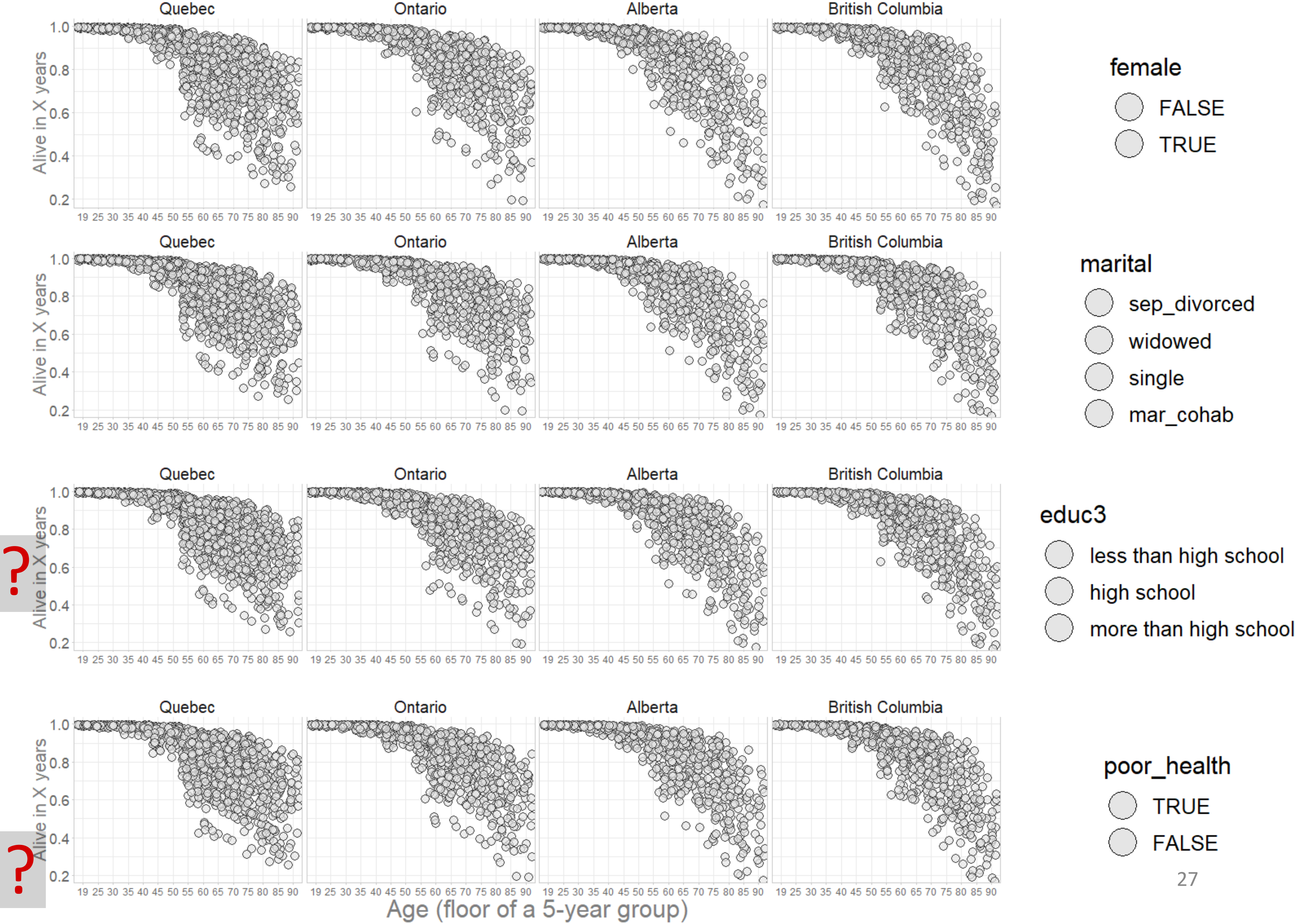
Substantially increased risk ?

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk ?



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

What levels of predictors are expected to affect mortality risk drastically?

Informed expectation

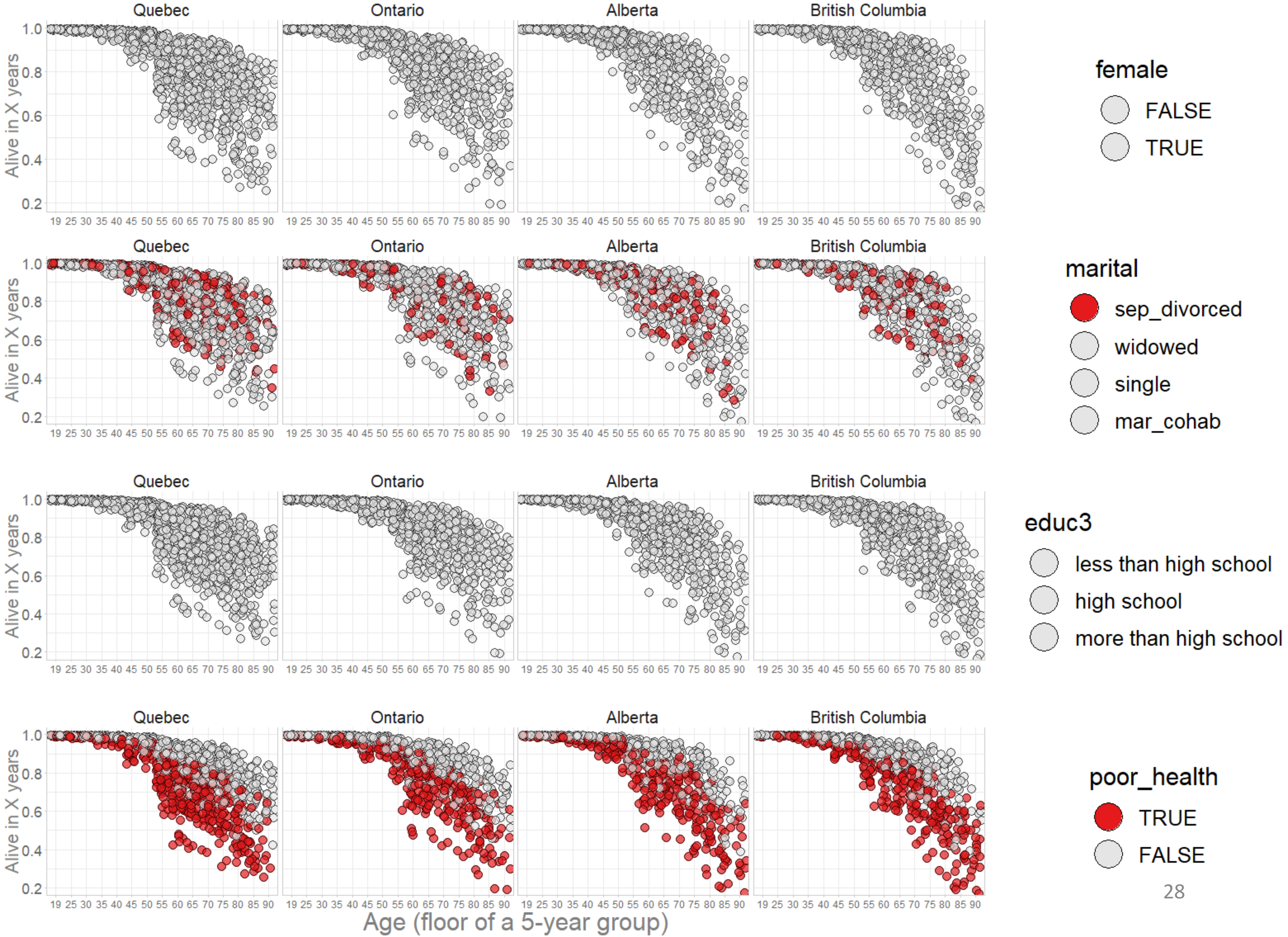
Substantially increased risk

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

QUESTION

What levels of predictors are expected to affect mortality risk drastically?

No “very bad” and it’s ok.

Informed expectation

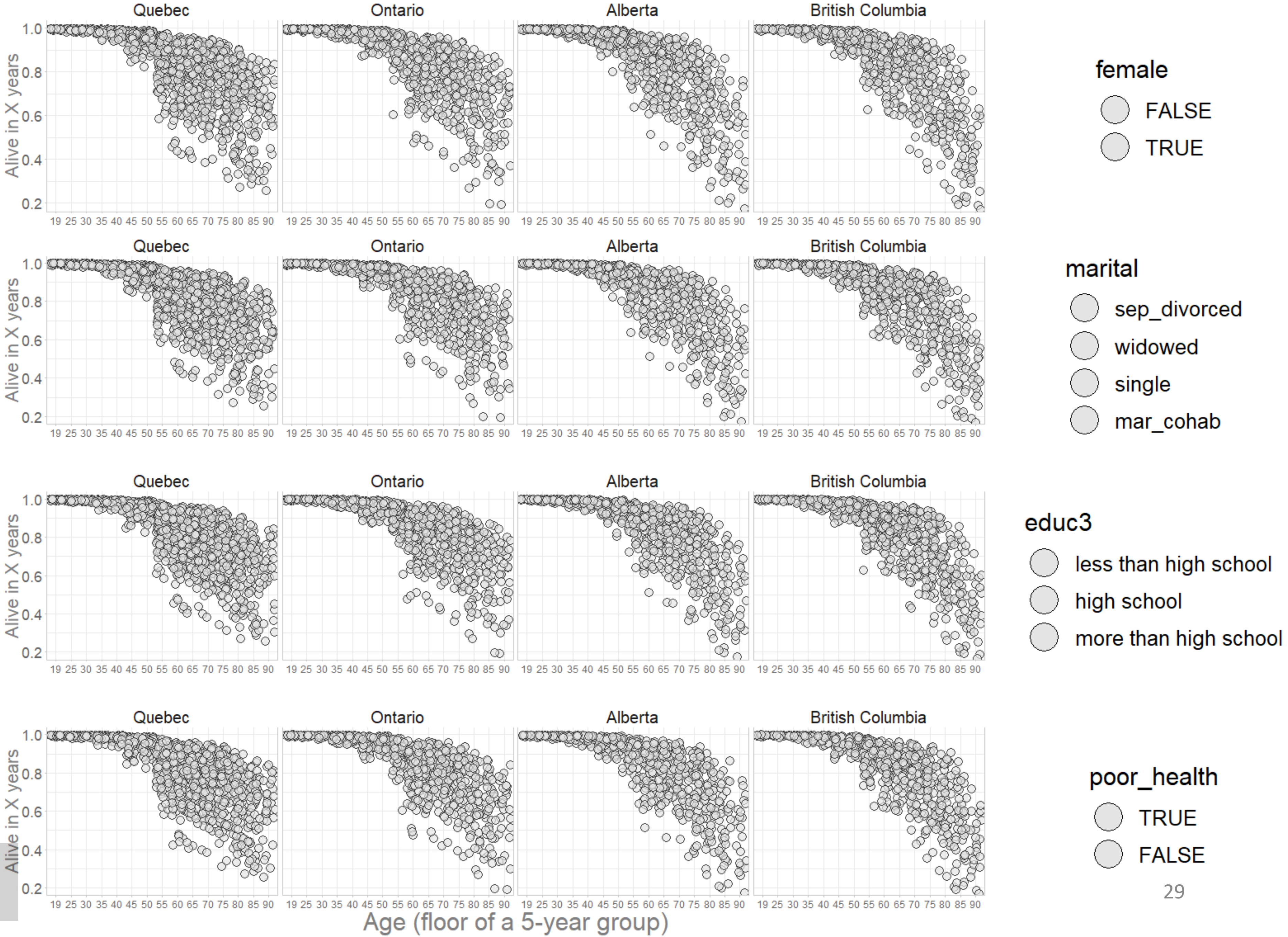
🟥 Substantially increased risk

🟧 Moderately increased risk

🟩 Reference group

🟦 Moderately decreased risk

🟪 Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

NOTICE

Plotting all colors at once may not be as informative as one would expect

May require too much tweaking to make useful

Informed expectation

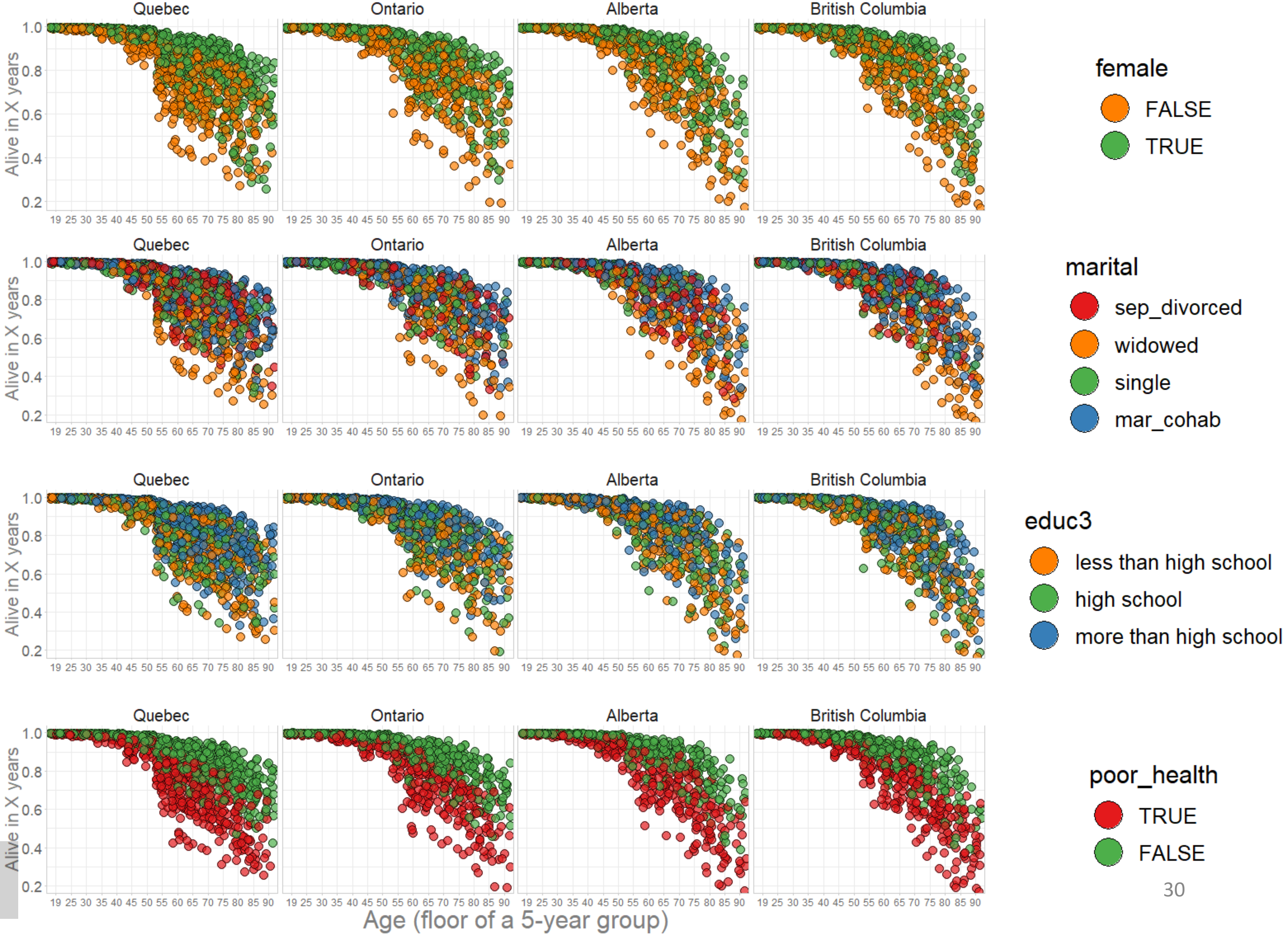
Substantially increased risk

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed

Informed expectation

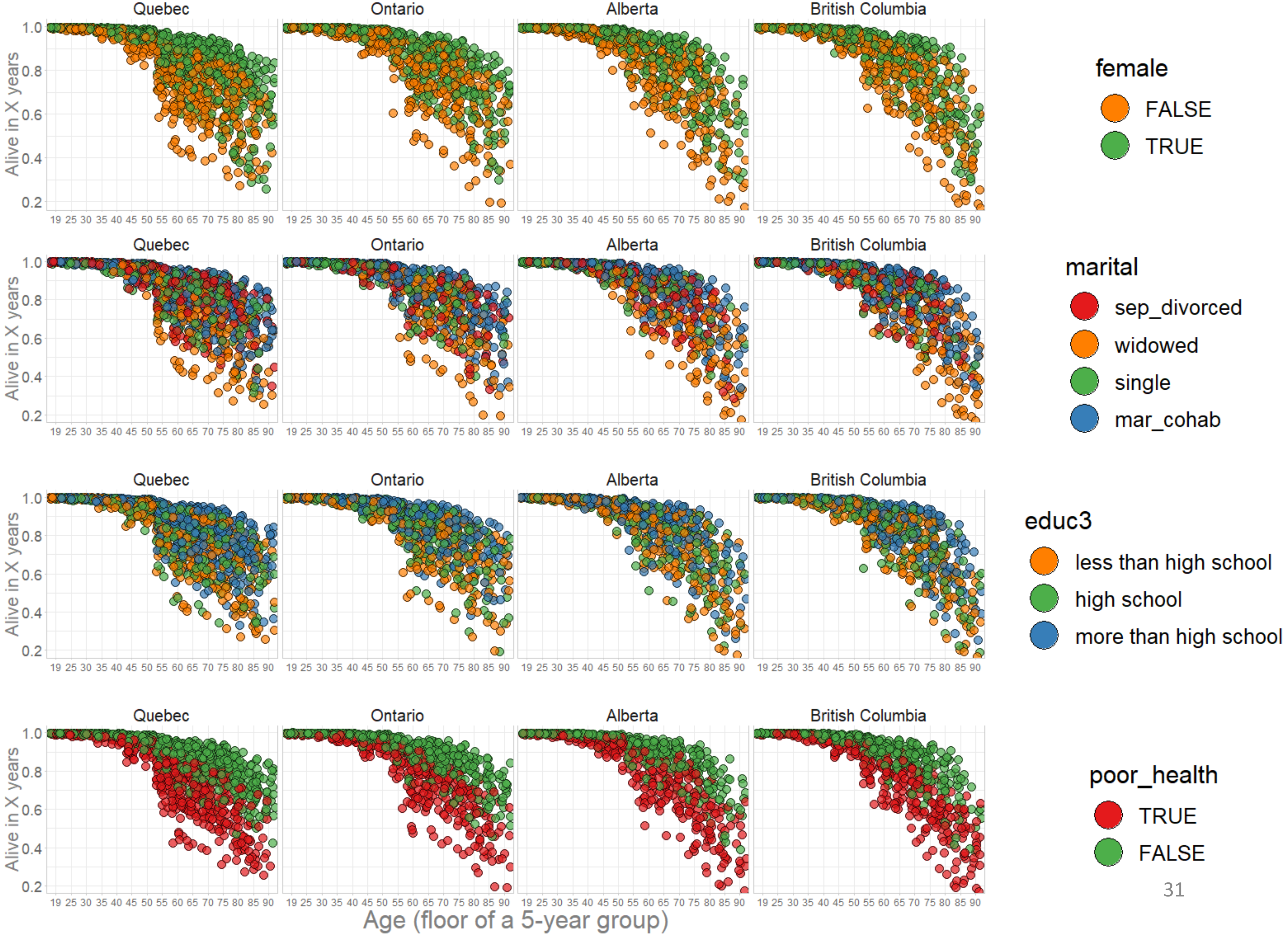
Substantially increased risk

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

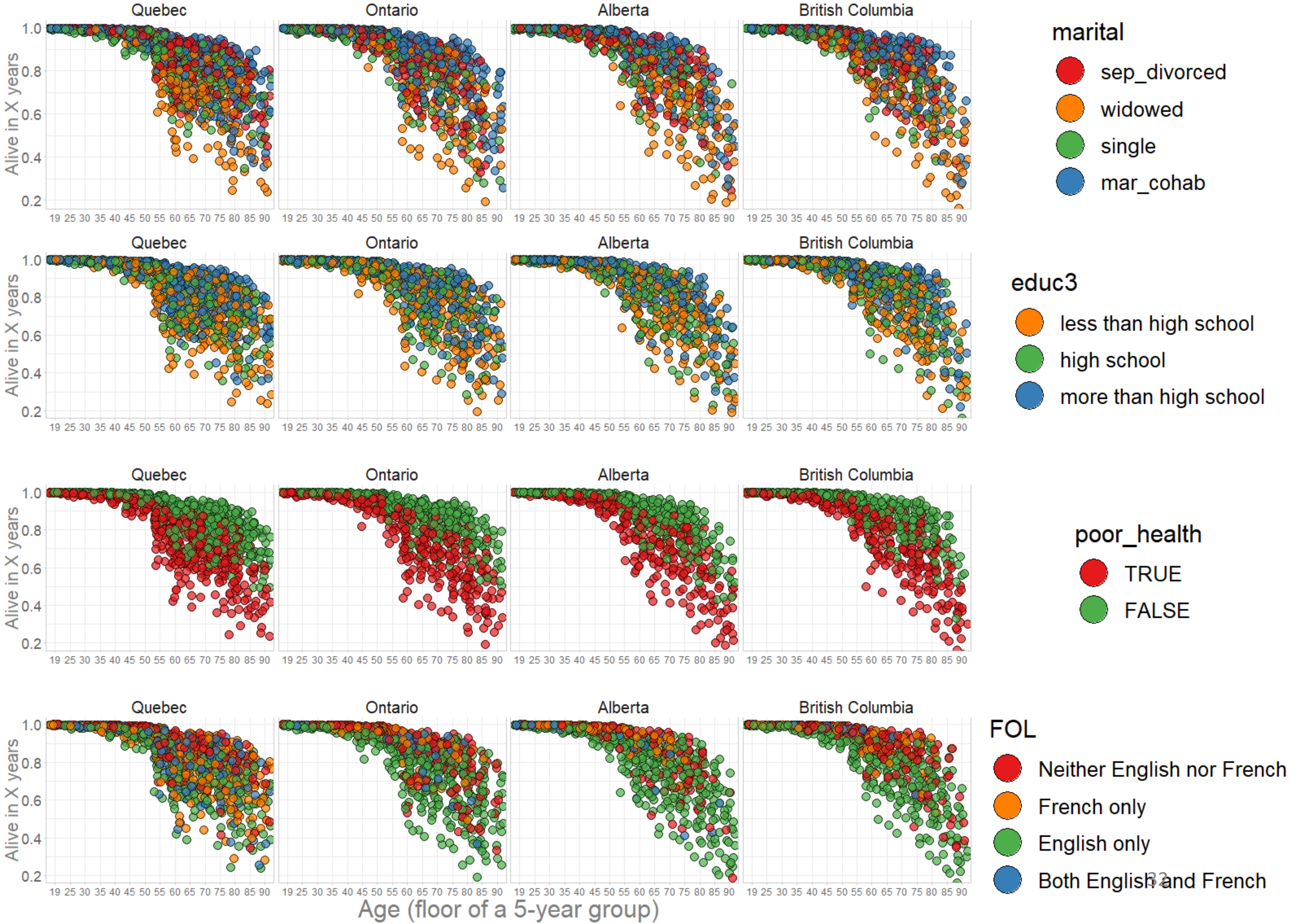
NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed

Informed expectation

- 🟥 Substantially increased risk
- 🟧 Moderately increased risk
- 🟩 Reference group
- 🟦 Moderately decreased risk
- 🟪 Substantially decreased risk



Shifting gears: IMPLEMENTATION

Questions to considered:

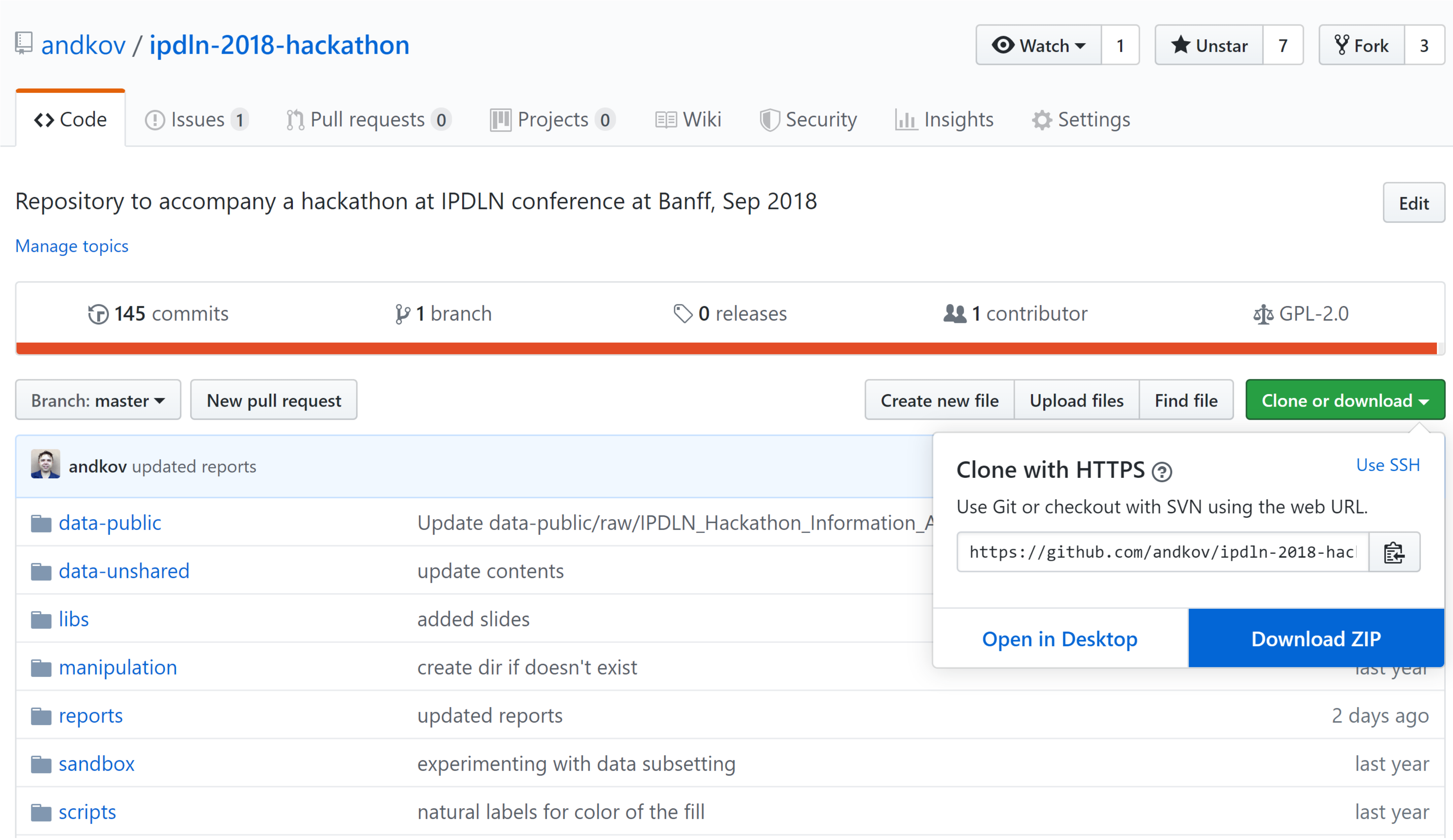
- How to organize files?
- What is a health degree of customization in graphs?
- Who are the future audience?
- How much of the story should be told?
- Do we expect to work on this in the future?
- How many people will be working on this?

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Autonomous phases: data cleaning, statistical modelling, graph production
1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

We will find these ideas implemented in this project

Clone to inspect the workflow



andkov / ipdln-2018-hackathon

Repository to accompany a hackathon at IPDLN conference at Banff, Sep 2018

andkov updated reports

data-public	Update data-public/raw/IPDLN_Hackathon_Information_A	
data-unshared	update contents	
libs	added slides	
manipulation	create dir if doesn't exist	last year
reports	updated reports	2 days ago
sandbox	experimenting with data subsetting	last year
scripts	natural labels for color of the fill	last year

Clone with HTTPS

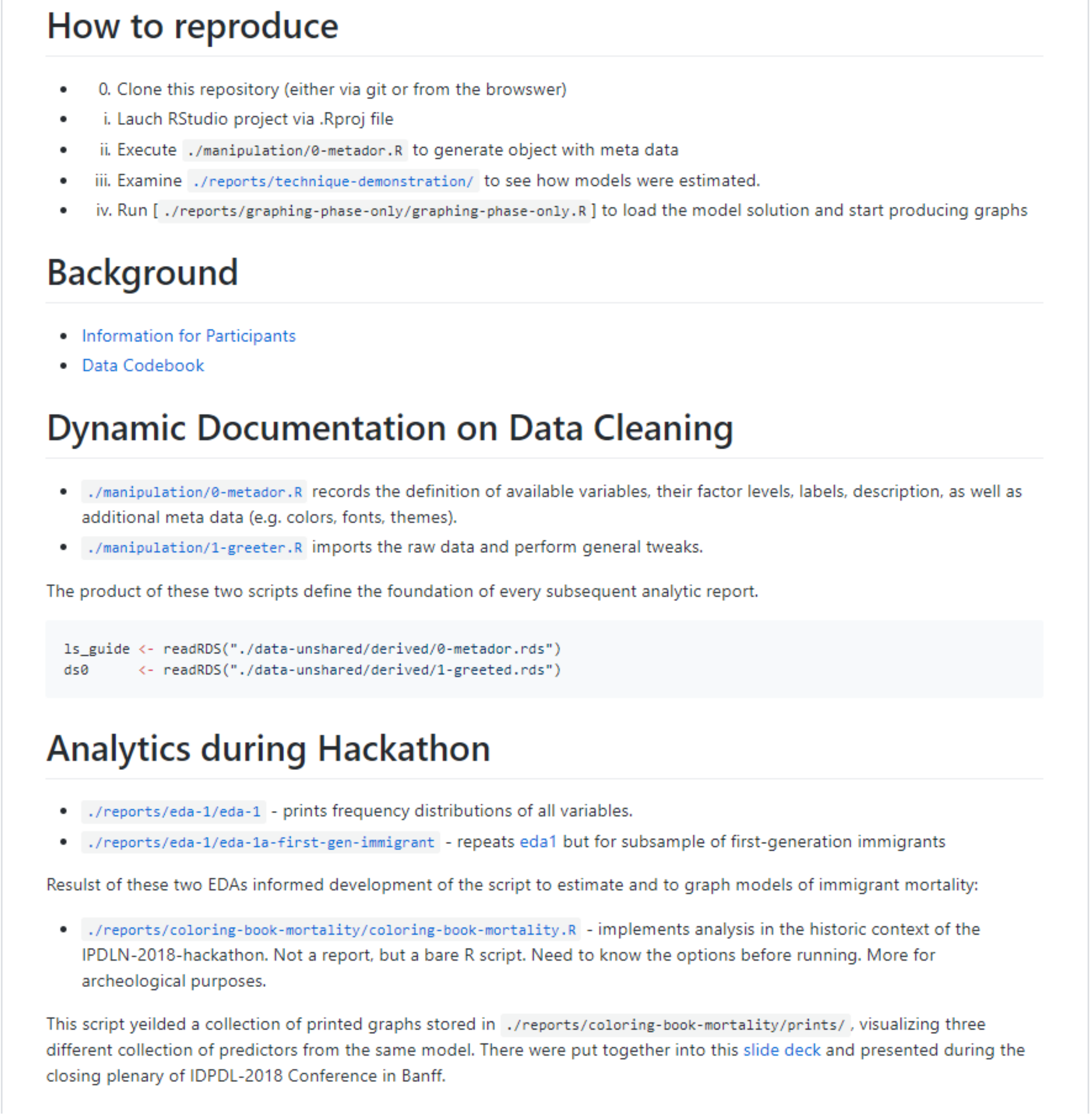
Use Git or checkout with SVN using the web URL.

```
https://github.com/andkov/ipdln-2018-hac
```

https://github.com/andkov/ipdln-2018-hackathon

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

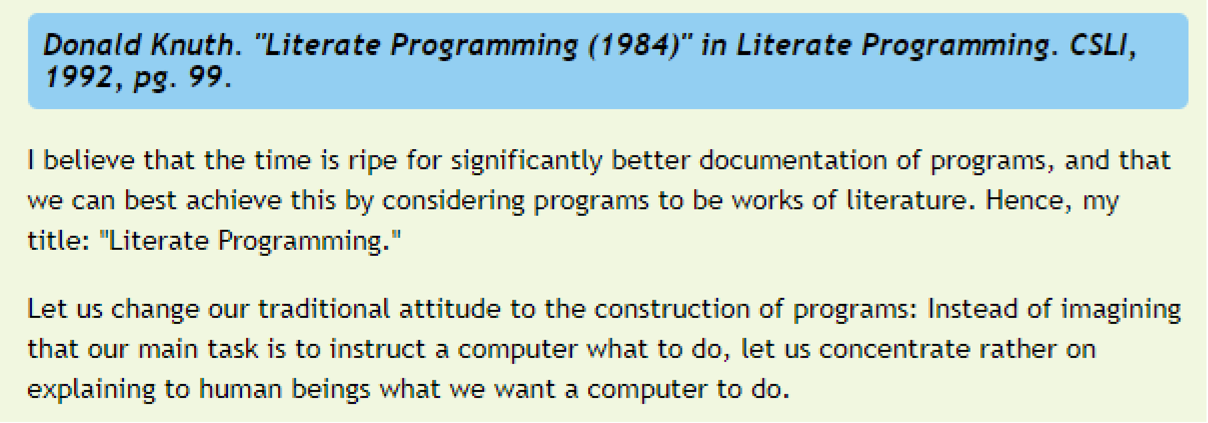
Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.



Donald Knuth. "Literate Programming (1984)" in Literate Programming. CSLI, 1992, pg. 99.

I believe that the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature. Hence, my title: "Literate Programming."

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.

Source: http://www.literateprogramming.com/

If you want to be a data scientist - expect to read scripts

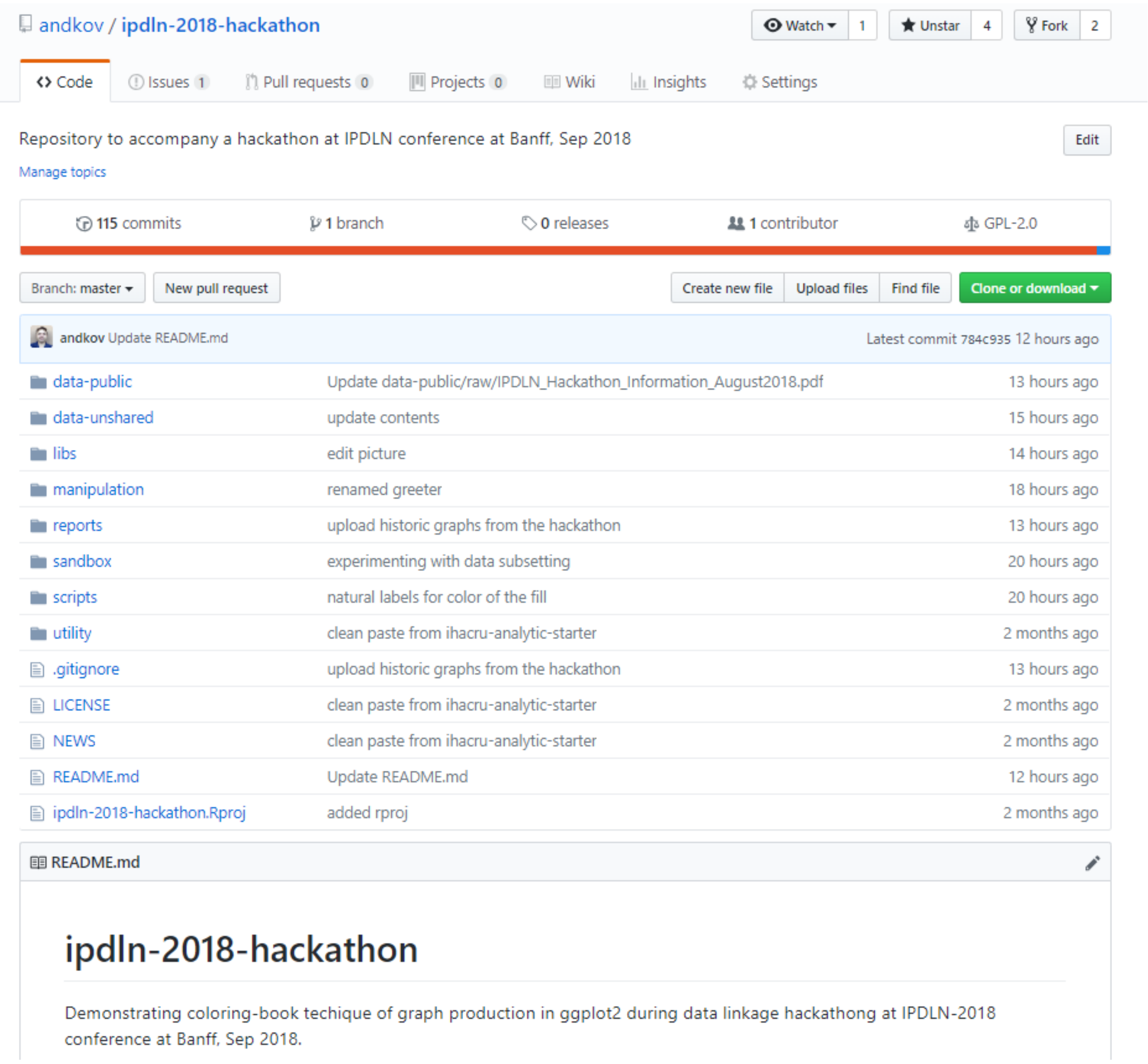
Main README should provide a map

https://github.com/andkov/ipdln-2018-hackathon/README.md

B. Workflow Highlights

1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects

Keep recognizable structure over projects



Repository to accompany a hackathon at IPDLN conference at Banff, Sep 2018

115 commits | 1 branch | 0 releases | 1 contributor | GPL-2.0

andkov Update README.md — Latest commit 784c935 12 hours ago

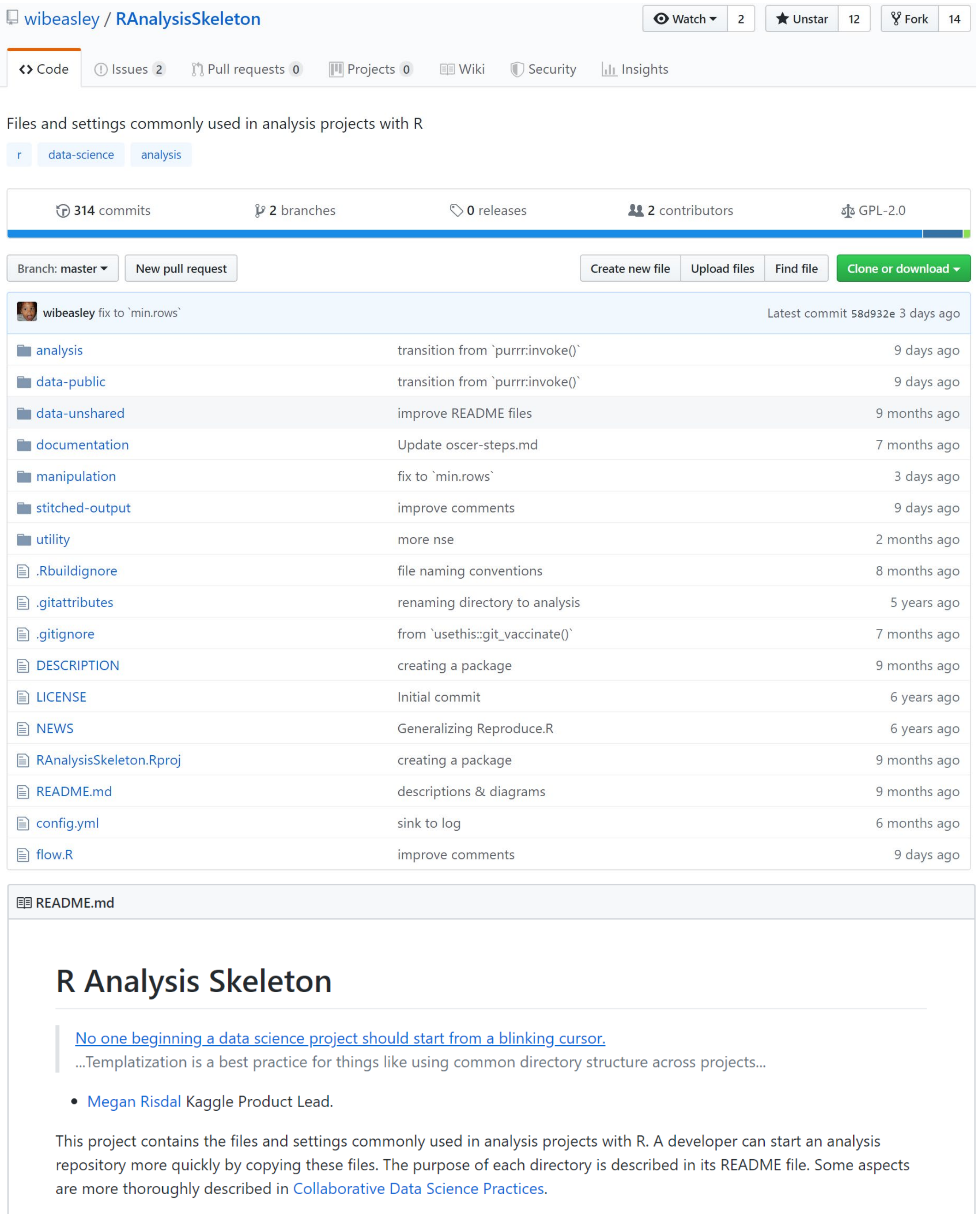
data-public	Update data-public/raw/IPDLN_Hackathon_Information_August2018.pdf	13 hours ago
data-unshared	update contents	15 hours ago
libs	edit picture	14 hours ago
manipulation	renamed greeter	18 hours ago
reports	upload historic graphs from the hackathon	13 hours ago
sandbox	experimenting with data subsetting	20 hours ago
scripts	natural labels for color of the fill	20 hours ago
utility	clean paste from ihacru-analytic-starter	2 months ago
.gitignore	upload historic graphs from the hackathon	13 hours ago
LICENSE	clean paste from ihacru-analytic-starter	2 months ago
NEWS	clean paste from ihacru-analytic-starter	2 months ago
README.md	Update README.md	12 hours ago
ipdln-2018-hackathon.Rproj	added rproj	2 months ago

README.md

ipdln-2018-hackathon

Demonstrating coloring-book techique of graph production in ggplot2 during data linkage hackathong at IPDLN-2018 conference at Banff, Sep 2018.

Notice structural similarities to RAnalysisSkeleton



Files and settings commonly used in analysis projects with R

314 commits | 2 branches | 0 releases | 2 contributors | GPL-2.0

wibeasley fix to `min.rows` — Latest commit 58d932e 3 days ago

analysis	transition from `purrr:invoke()`	9 days ago
data-public	transition from `purrr:invoke()`	9 days ago
data-unshared	improve README files	9 months ago
documentation	Update oscer-steps.md	7 months ago
manipulation	fix to `min.rows`	3 days ago
stitched-output	improve comments	9 days ago
utility	more nse	2 months ago
.Rbuildignore	file naming conventions	8 months ago
.gitattributes	renaming directory to analysis	5 years ago
.gitignore	from `usethis::git_vaccinate()`	7 months ago
DESCRIPTION	creating a package	9 months ago
LICENSE	Initial commit	6 years ago
NEWS	Generalizing Reproduce.R	6 years ago
RAnalysisSkeleton.Rproj	creating a package	9 months ago
README.md	descriptions & diagrams	9 months ago
config.yml	sink to log	6 months ago
flow.R	improve comments	9 days ago

README.md

R Analysis Skeleton

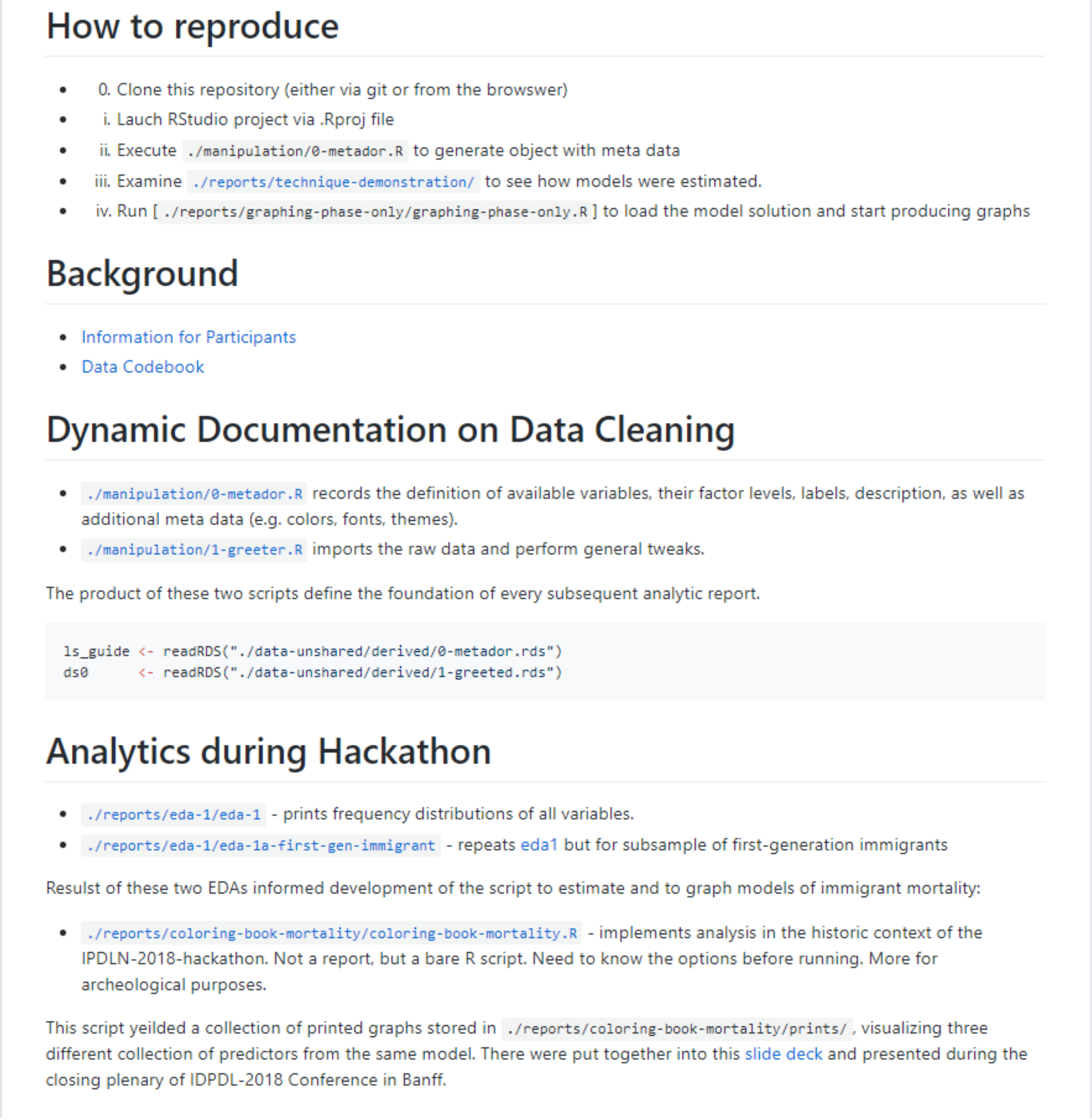
> No one beginning a data science project should start from a blinking cursor.
> ...Templatization is a best practice for things like using common directory structure across projects...

- Megan Risdal Kaggle Product Lead.

This project contains the files and settings commonly used in analysis projects with R. A developer can start an analysis repository more quickly by copying these files. The purpose of each directory is described in its README file. Some aspects are more thoroughly described in Collaborative Data Science Practices.

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

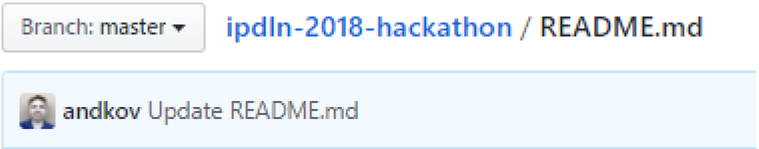
Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.



ipdln-2018-hackathon / README.md

andkov Update README.md

Try to keep tasks separate:

- Data cleaning
- Statistical modeling
- Graph production

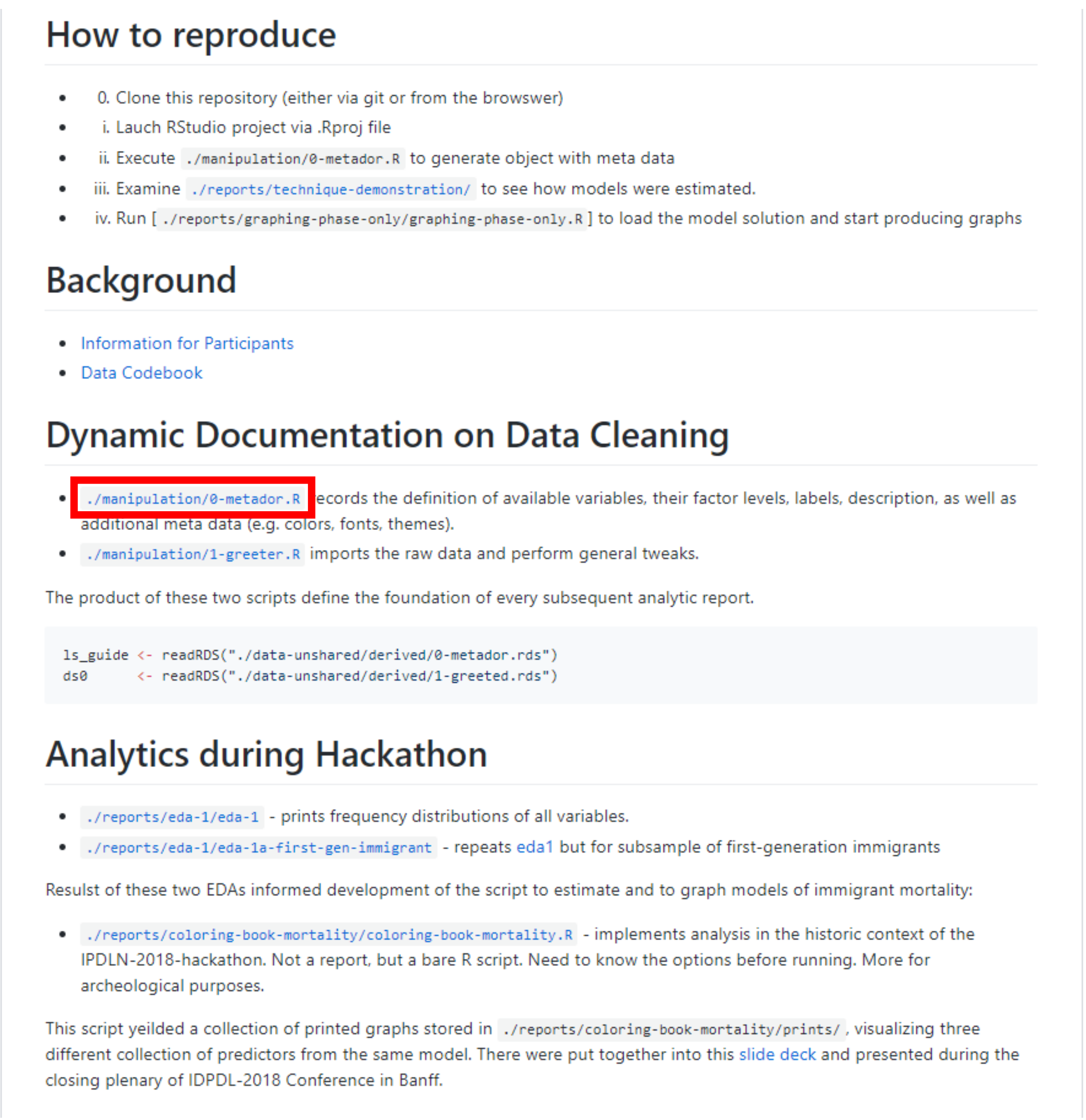
Tasks are narratives to be told

Here are some examples

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

Screenshots of linked dynamic document



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` ecords the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

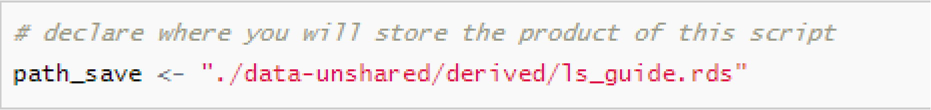
Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

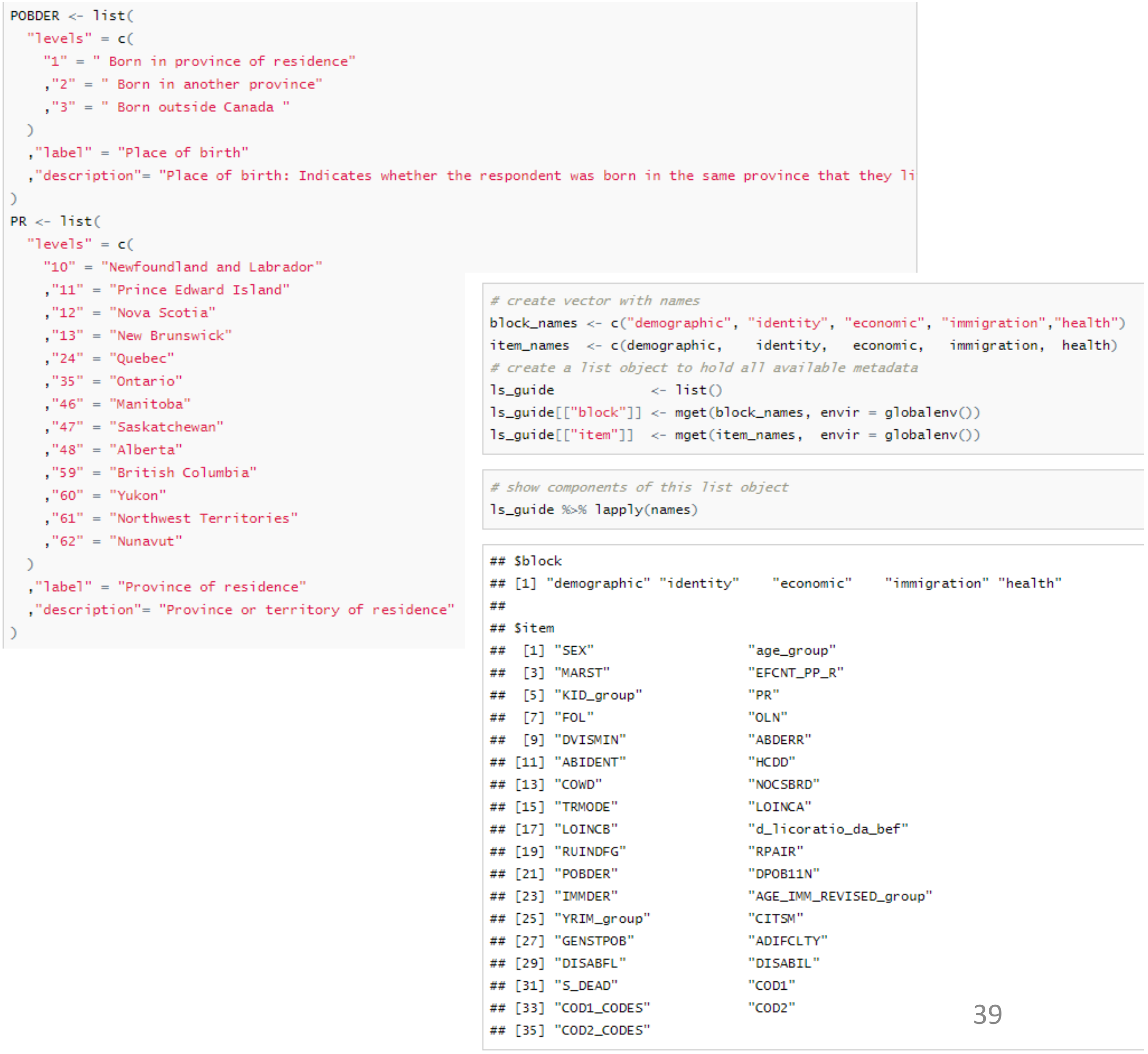
Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.



```
# declare where you will store the product of this script
path_save <- "./data-unshared/derived/ls_guide.rds"
```



```
POBDER <- list(
  "levels" = c(
    "1" = " Born in province of residence"
    ,"2" = " Born in another province"
    ,"3" = " Born outside Canada "
  )
  ,"label" = "Place of birth"
  ,"description"= "Place of birth: Indicates whether the respondent was born in the same province that they li
)
PR <- list(
  "levels" = c(
    "10" = "Newfoundland and Labrador"
    ,"11" = "Prince Edward Island"
    ,"12" = "Nova Scotia"
    ,"13" = "New Brunswick"
    ,"24" = "Quebec"
    ,"35" = "Ontario"
    ,"46" = "Manitoba"
    ,"47" = "Saskatchewan"
    ,"48" = "Alberta"
    ,"59" = "British Columbia"
    ,"60" = "Yukon"
    ,"61" = "Northwest Territories"
    ,"62" = "Nunavut"
  )
  ,"label" = "Province of residence"
  ,"description"= "Province or territory of residence"
)
```

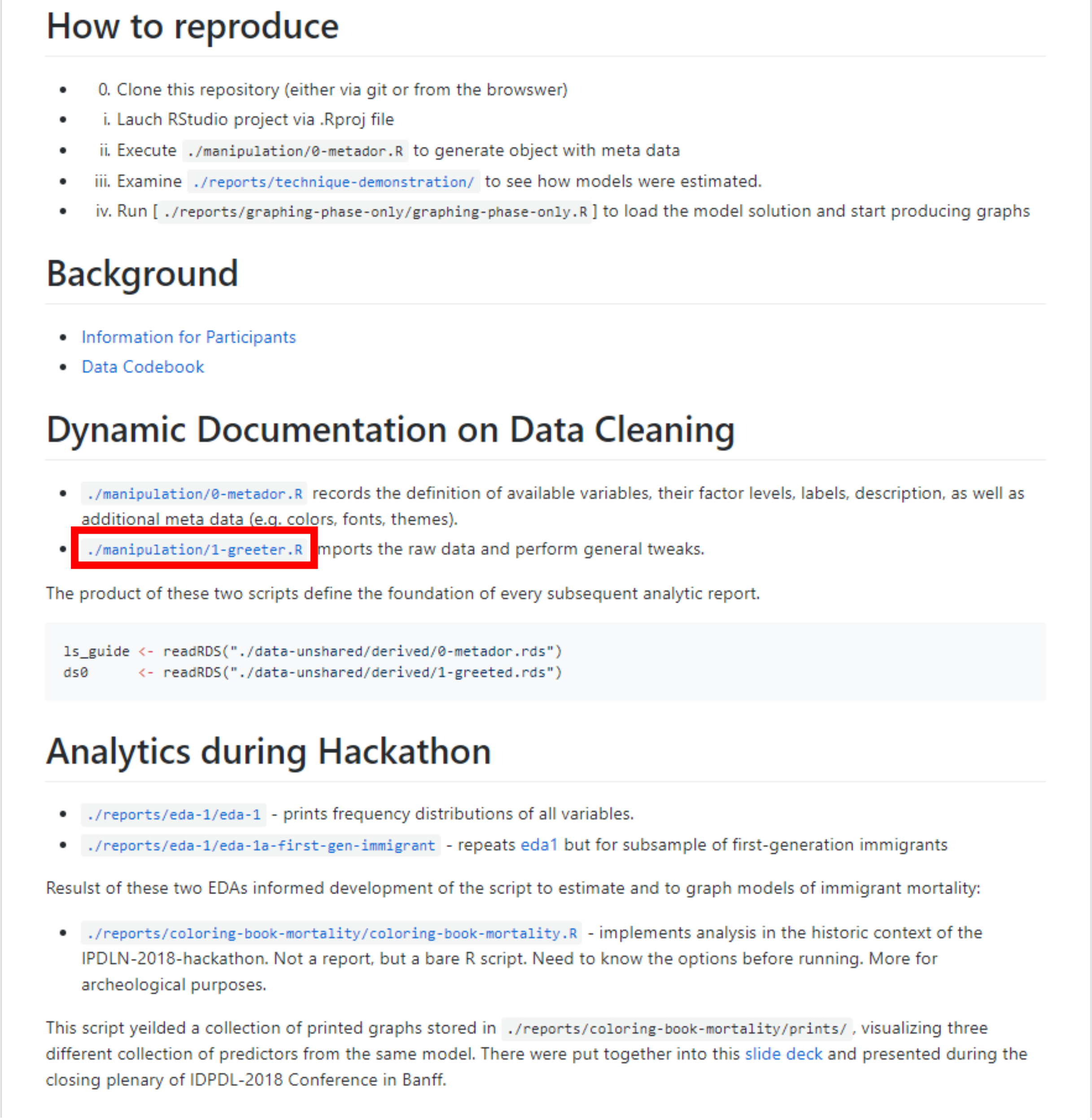
```
# create vector with names
block_names <- c("demographic", "identity", "economic", "immigration","health")
item_names  <- c(demographic,    identity,   economic,   immigration,  health)
# create a list object to hold all available metadata
ls_guide            <- list()
ls_guide[["block"]] <- mget(block_names, envir = globalenv())
ls_guide[["item"]]  <- mget(item_names,  envir = globalenv())
```

```
# show components of this list object
ls_guide %>% lapply(names)
```

```
## $block
## [1] "demographic" "identity"    "economic"    "immigration" "health"
##
## $item
##  [1] "SEX"                 "age_group"
##  [3] "MARST"               "EFCNT_PP_R"
##  [5] "KID_group"           "PR"
##  [7] "FOL"                 "OLN"
##  [9] "DVISMIN"             "ABDERR"
## [11] "ABIDENT"             "HCDD"
## [13] "COWD"                "NOCSBRD"
## [15] "TRMODE"              "LOINCA"
## [17] "LOINCB"              "d_licoratio_da_bef"
## [19] "RUINDFG"             "RPAIR"
## [21] "POBDER"              "DPOB11N"
## [23] "IMMDER"              "AGE_IMM_REVISED_group"
## [25] "YRIM_group"          "CITSM"
## [27] "GENSTPOB"            "ADIFCLTY"
## [29] "DISABFL"             "DISABIL"
## [31] "S_DEAD"              "COD1"
## [33] "COD1_CODES"          "COD2"
## [35] "COD2_CODES"
```

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

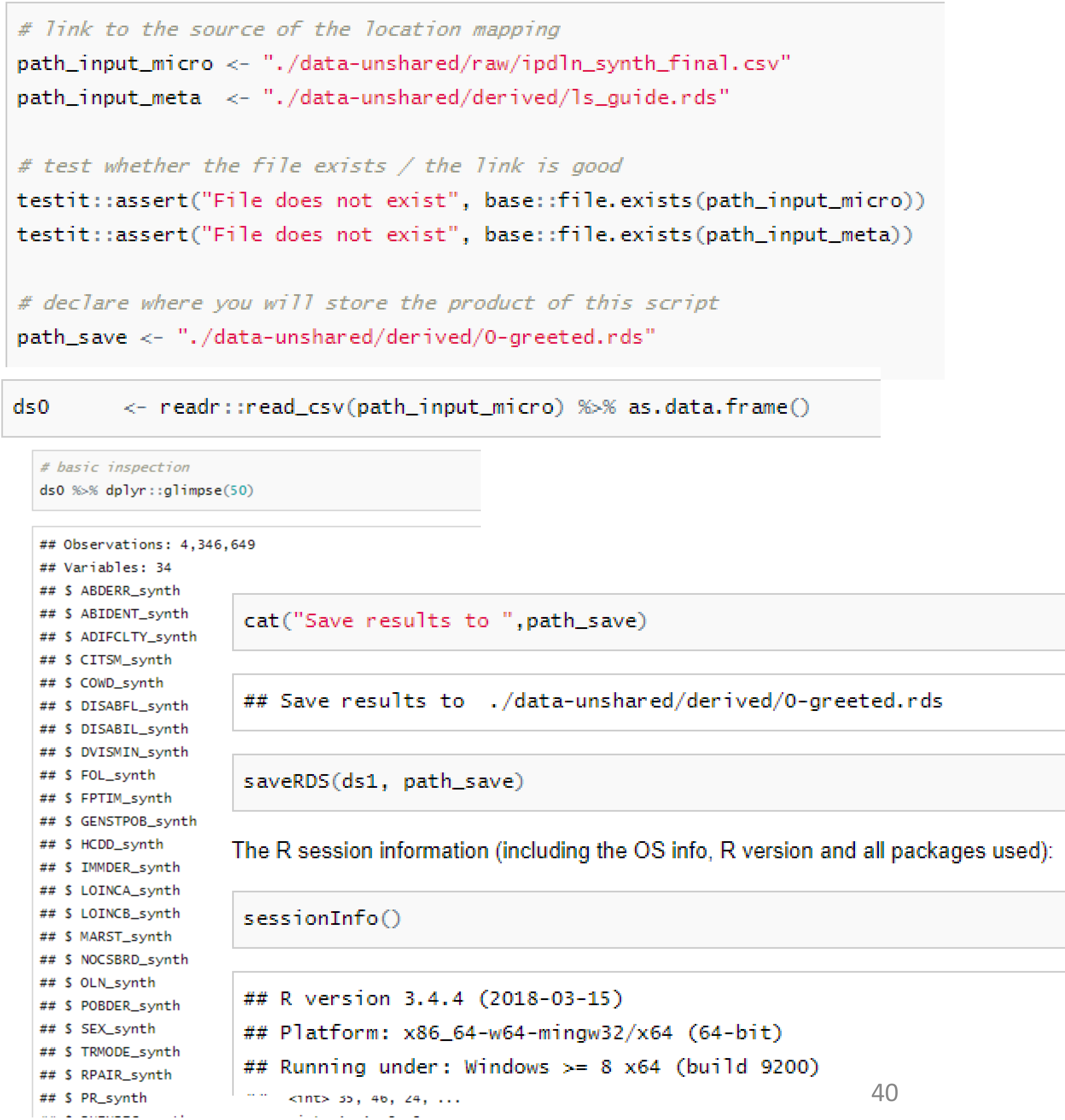
- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document



```
# link to the source of the location mapping
path_input_micro <- "./data-unshared/raw/ipdln_synth_final.csv"
path_input_meta  <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
testit::assert("File does not exist", base::file.exists(path_input_micro))
testit::assert("File does not exist", base::file.exists(path_input_meta))

# declare where you will store the product of this script
path_save <- "./data-unshared/derived/0-greeted.rds"
```

```
ds0      <- readr::read_csv(path_input_micro) %>% as.data.frame()
```

```
# basic inspection
ds0 %>% dplyr::glimpse(50)
```

```
## Observations: 4,346,649
## Variables: 34
## $ ABDERR_synth
## $ ABIDENT_synth
## $ ADIFCLTY_synth
## $ CITSM_synth
## $ COWD_synth
## $ DISABFL_synth
## $ DISABIL_synth
## $ DVISMIN_synth
## $ FOL_synth
## $ FPTIM_synth
## $ GENSTPOB_synth
## $ HCDD_synth
## $ IMMDER_synth
## $ LOINCA_synth
## $ LOINCB_synth
## $ MARST_synth
## $ NOCSBRD_synth
## $ OLN_synth
## $ POBDER_synth
## $ SEX_synth
## $ TRMODE_synth
## $ RPAIR_synth
## $ PR_synth
```

```
cat("Save results to ",path_save)
```

```
## Save results to  ./data-unshared/derived/0-greeted.rds
```

```
saveRDS(ds1, path_save)
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()
```

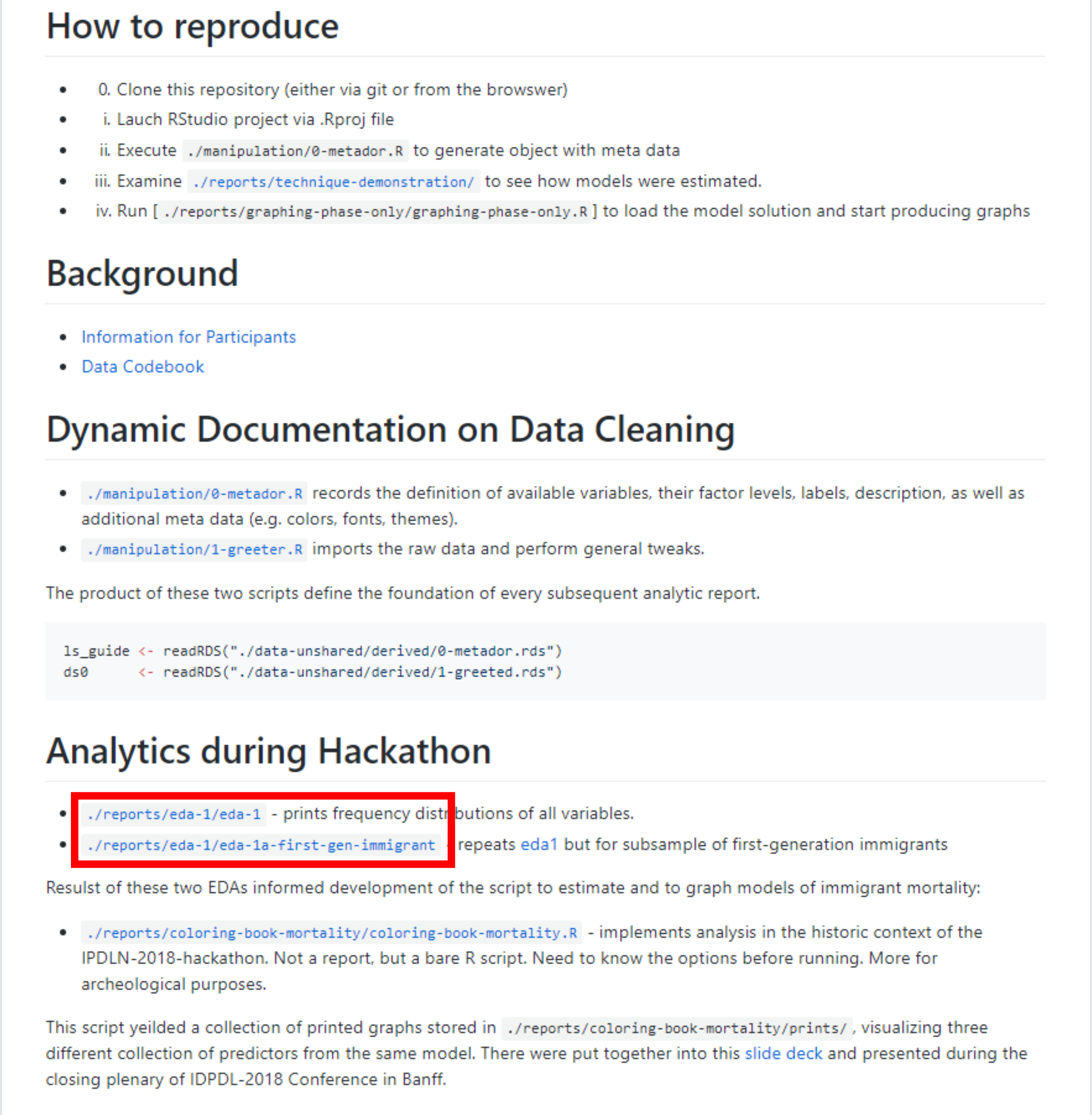
```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows >= 8 x64 (build 9200)
```

```
<int> 55, 46, 24, ...
```

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

Screenshots of linked dynamic document



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

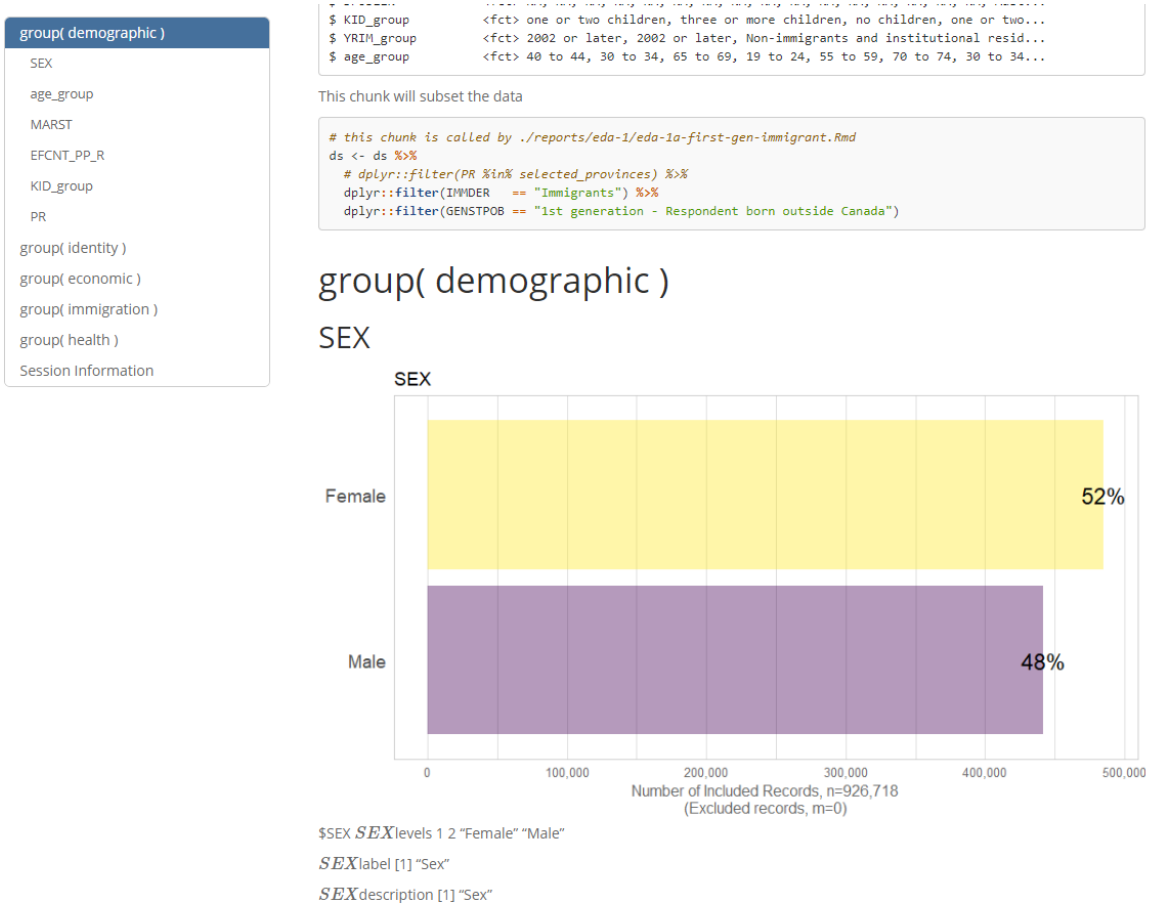
Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.



```
$ KID_group       <fct> one or two children, three or more children, no children, one or two...
$ YRIM_group      <fct> 2002 or later, 2002 or later, Non-immigrants and institutional resid...
$ age_group       <fct> 40 to 44, 30 to 34, 65 to 69, 19 to 24, 55 to 59, 70 to 74, 30 to 34...
```

This chunk will subset the data

```
# this chunk is called by ./reports/eda-1/eda-1a-first-gen-immigrant.Rmd
ds <- ds %>%
  # dplyr::filter(PR %in% selected_provinces) %>%
  dplyr::filter(IMMDER   == "Immigrants") %>%
  dplyr::filter(GENSTPOB == "1st generation - Respondent born outside Canada")
```

group(demographic)

SEX

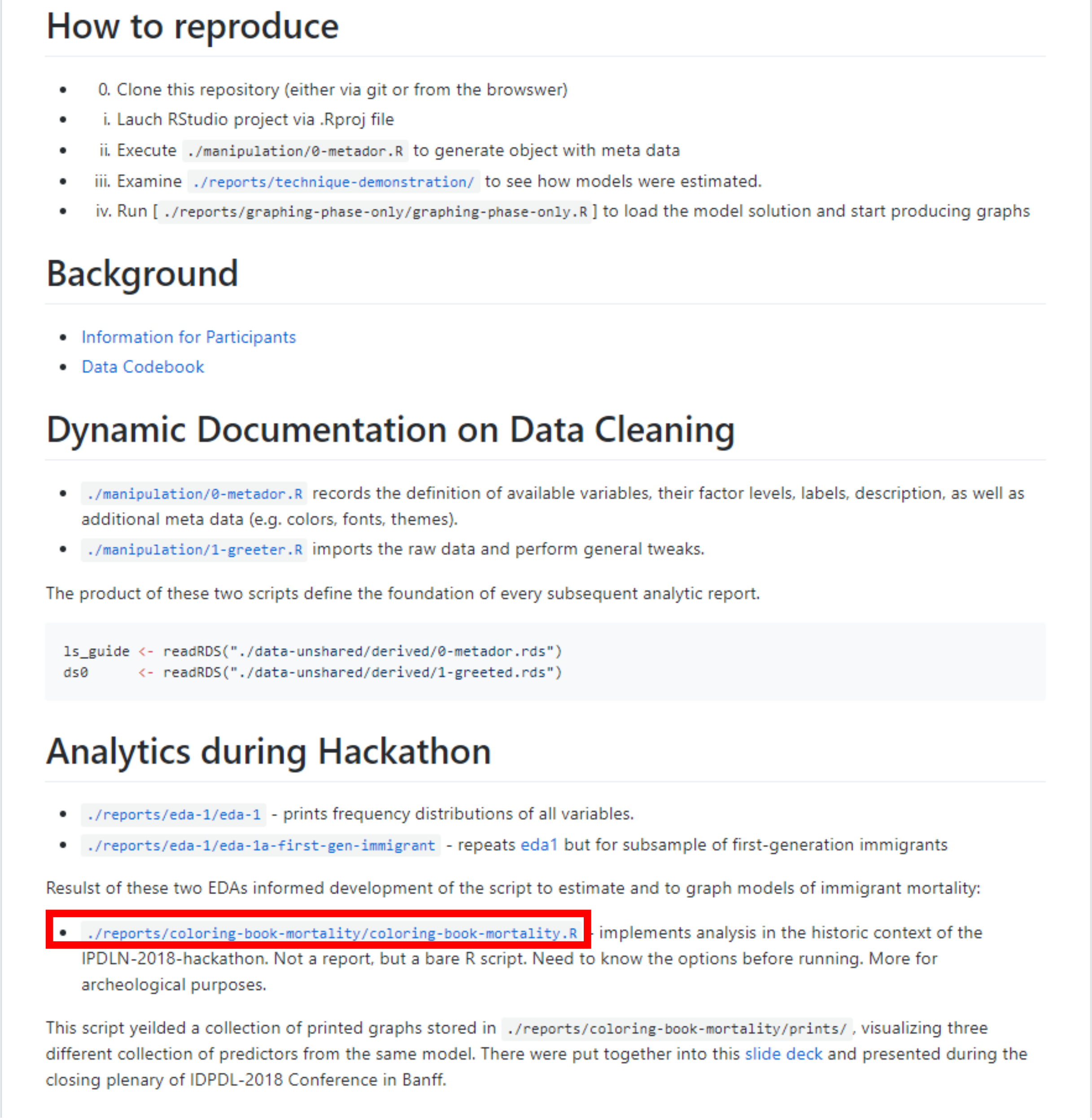
$SEX SEXlevels 1 2 “Female” “Male”

SEXlabel [1] “Sex”

SEXdescription [1] “Sex”

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production



How to reproduce

- 0. Clone this repository (either via git or from the browswer)
- i. Lauch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run [`./reports/graphing-phase-only/graphing-phase-only.R`] to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

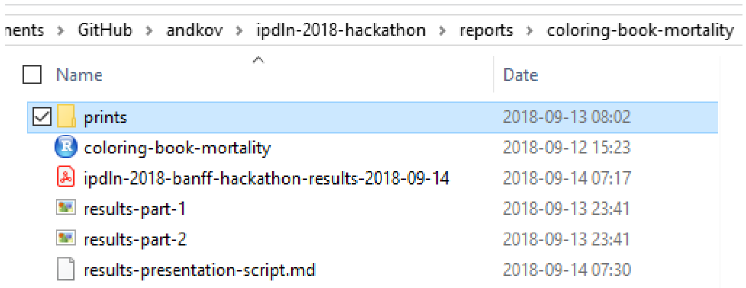
- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats eda1 but for subsample of first-generation immigrants

Resulst of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

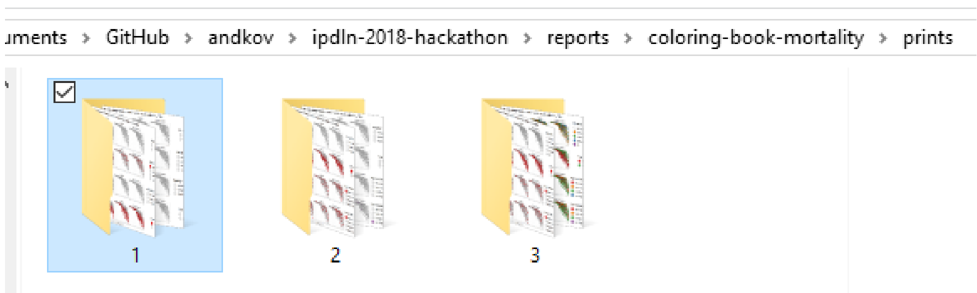
This script yeilded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this slide deck and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of project repository

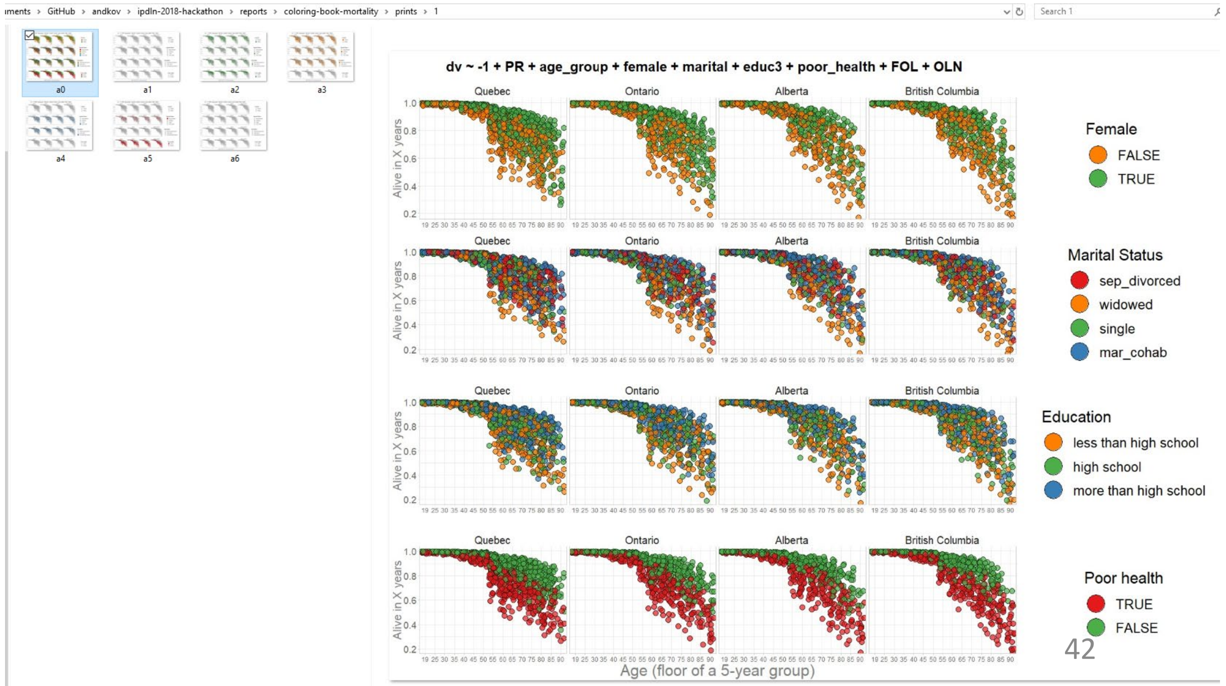


ıents › GitHub › andkov › ipdln-2018-hackathon › reports › coloring-book-mortality

□ Name	Date
☑ prints	2018-09-13 08:02
coloring-book-mortality	2018-09-12 15:23
ipdln-2018-banff-hackathon-results-2018-09-14	2018-09-14 07:17
results-part-1	2018-09-13 23:41
results-part-2	2018-09-13 23:41
results-presentation-script.md	2018-09-14 07:30



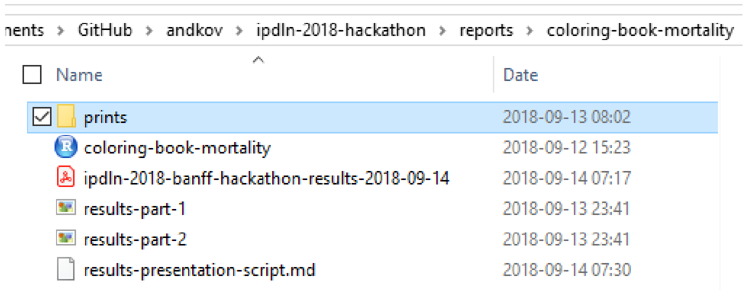
ıments › GitHub › andkov › ipdln-2018-hackathon › reports › coloring-book-mortality › prints
1
2
3



B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

Screenshots of project repository

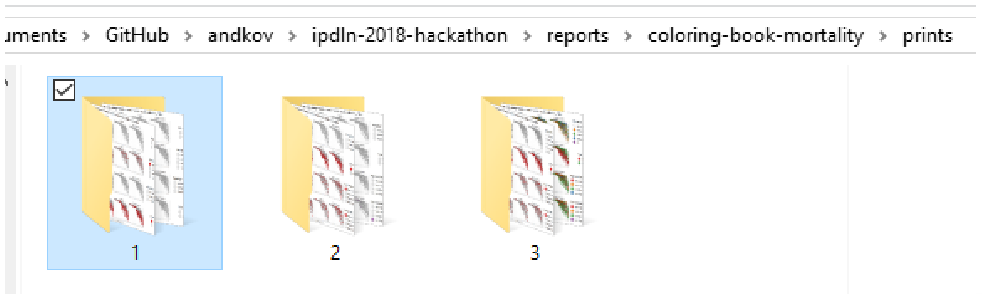


ıents › GitHub › andkov › ipdln-2018-hackathon › reports › coloring-book-mortality

□ Name	Date
☑ prints	2018-09-13 08:02
coloring-book-mortality	2018-09-12 15:23
ipdln-2018-banff-hackathon-results-2018-09-14	2018-09-14 07:17
results-part-1	2018-09-13 23:41
results-part-2	2018-09-13 23:41
results-presentation-script.md	2018-09-14 07:30

./reports/coloring-book-mortality/

Fails to separate modeling, graphing, and reporting

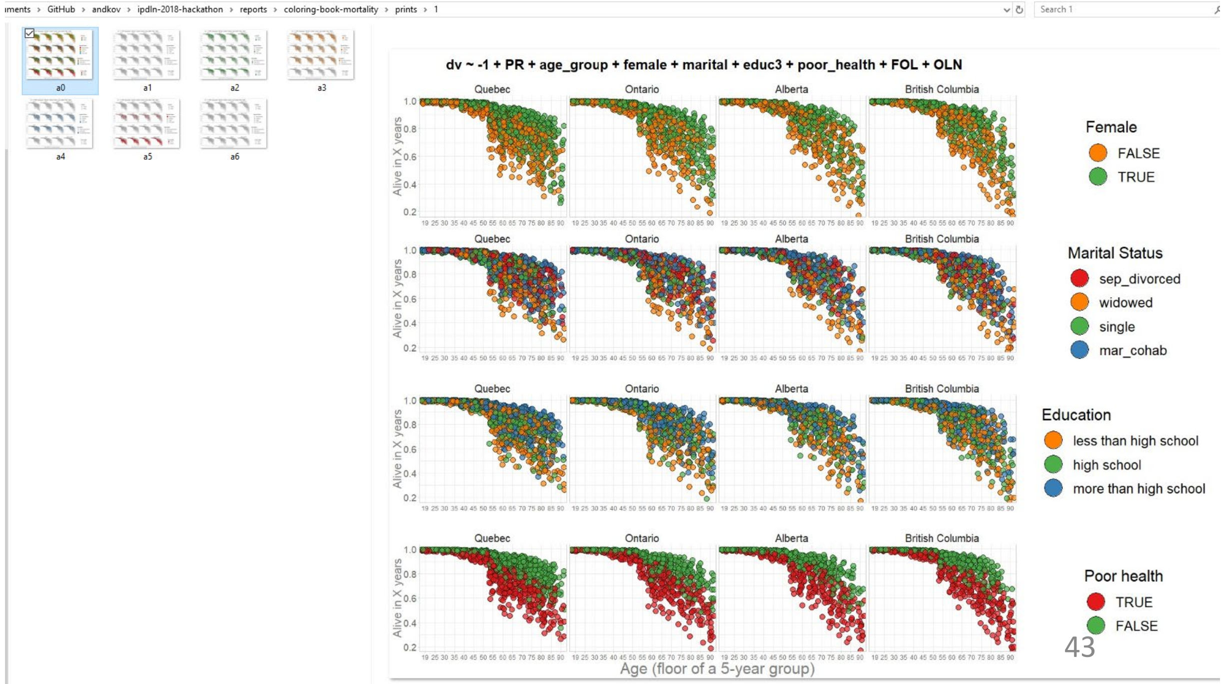


ıments › GitHub › andkov › ipdln-2018-hackathon › reports › coloring-book-mortality › prints

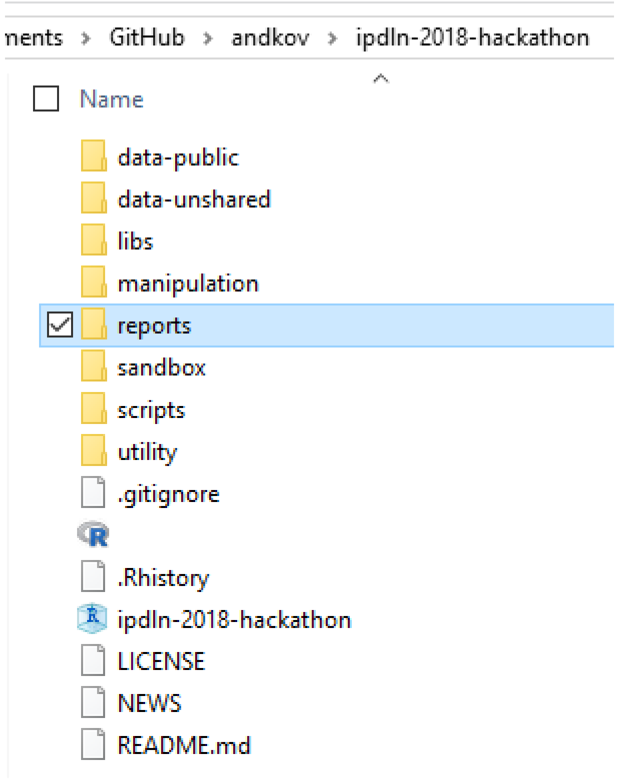
☑ 1

2

3



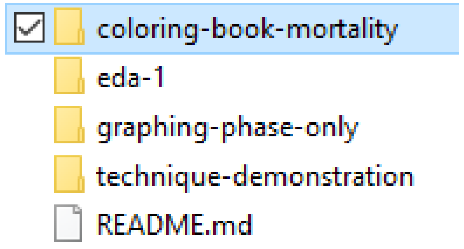
ıments › GitHub › andkov › ipdln-2018-hackathon › reports › coloring-book-mortality › prints › 1



ıents › GitHub › andkov › ipdln-2018-hackathon

□ Name

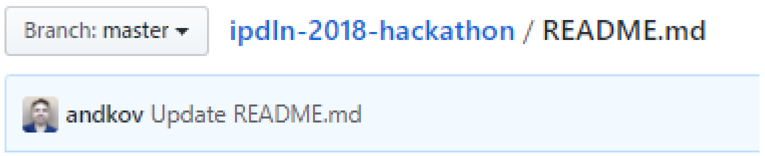
- data-public
- data-unshared
- libs
- manipulation
- ☑ reports
- sandbox
- scripts
- utility
- .gitignore
- .Rhistory
- ipdln-2018-hackathon
- LICENSE
- NEWS
- README.md



- ☑ coloring-book-mortality
- eda-1
- graphing-phase-only
- technique-demonstration
- README.md

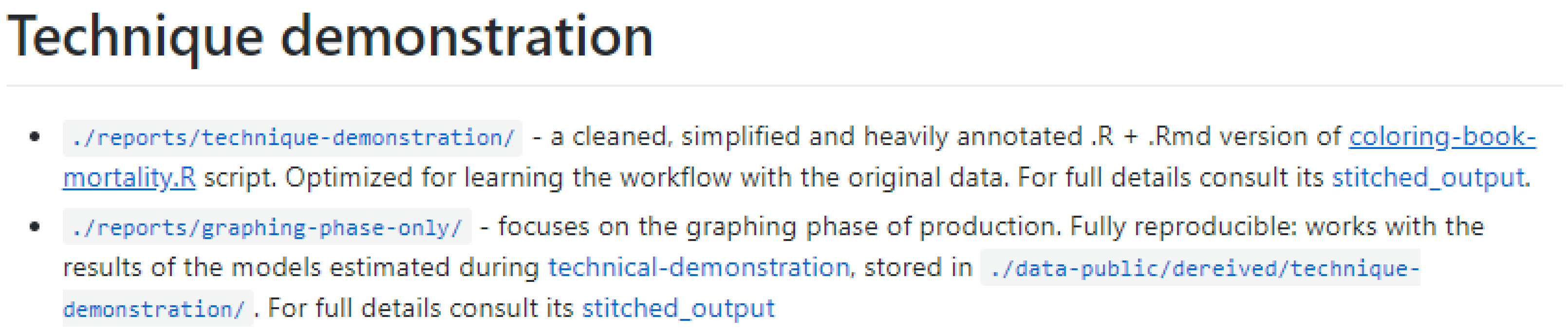
B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html



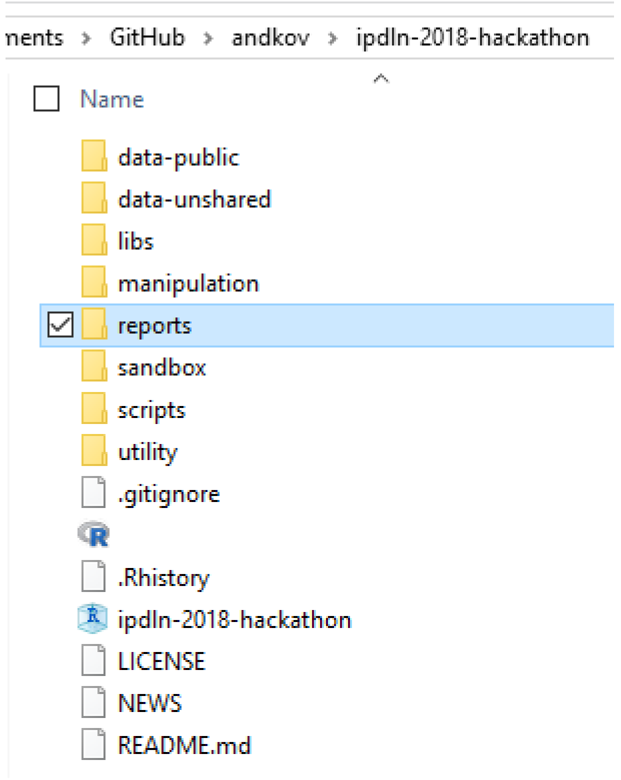
Branch: master ▾ ipdln-2018-hackathon / README.md

andkov Update README.md



Technique demonstration

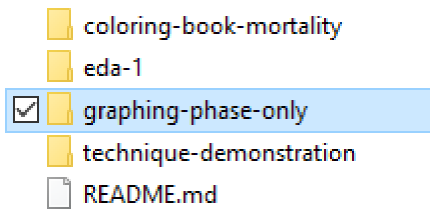
- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of coloring-book-mortality.R script. Optimized for learning the workflow with the original data. For full details consult its stitched_output.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during technical-demonstration, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its stitched_output



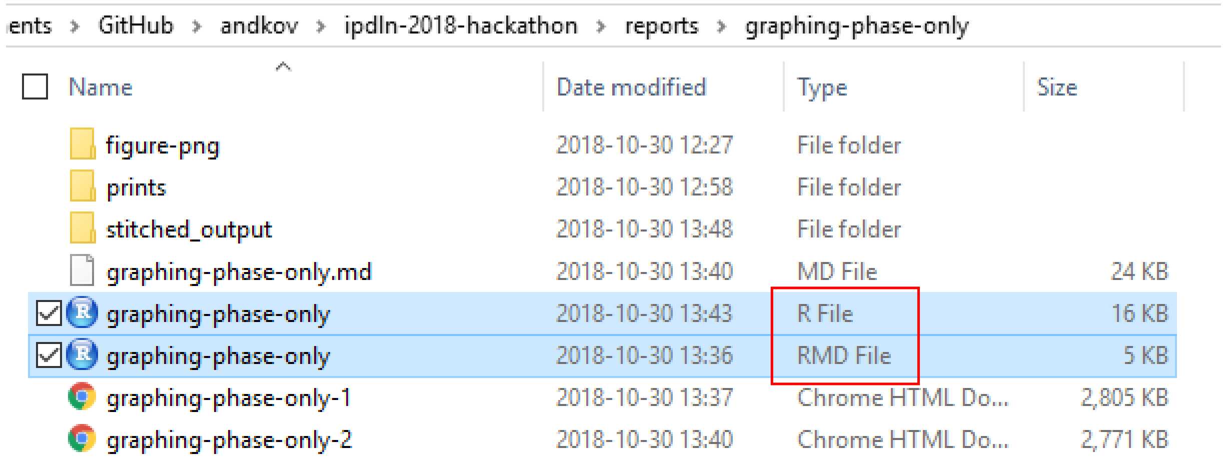
nents › GitHub › andkov › ipdln-2018-hackathon

□ Name

- data-public
- data-unshared
- libs
- manipulation
- ☑ reports
- sandbox
- scripts
- utility
- .gitignore
- .Rhistory
- ipdln-2018-hackathon
- LICENSE
- NEWS
- README.md



- coloring-book-mortality
- eda-1
- ☑ graphing-phase-only
- technique-demonstration
- README.md



ents › GitHub › andkov › ipdln-2018-hackathon › reports › graphing-phase-only

□	Name	Date modified	Type	Size
	figure-png	2018-10-30 12:27	File folder	
	prints	2018-10-30 12:58	File folder	
	stitched_output	2018-10-30 13:48	File folder	
	graphing-phase-only.md	2018-10-30 13:40	MD File	24 KB
☑	graphing-phase-only	2018-10-30 13:43	R File	16 KB
☑	graphing-phase-only	2018-10-30 13:36	RMD File	5 KB
	graphing-phase-only-1	2018-10-30 13:37	Chrome HTML Do...	2,805 KB
	graphing-phase-only-2	2018-10-30 13:40	Chrome HTML Do...	2,771 KB

.R
stores analysis
(what really happens)

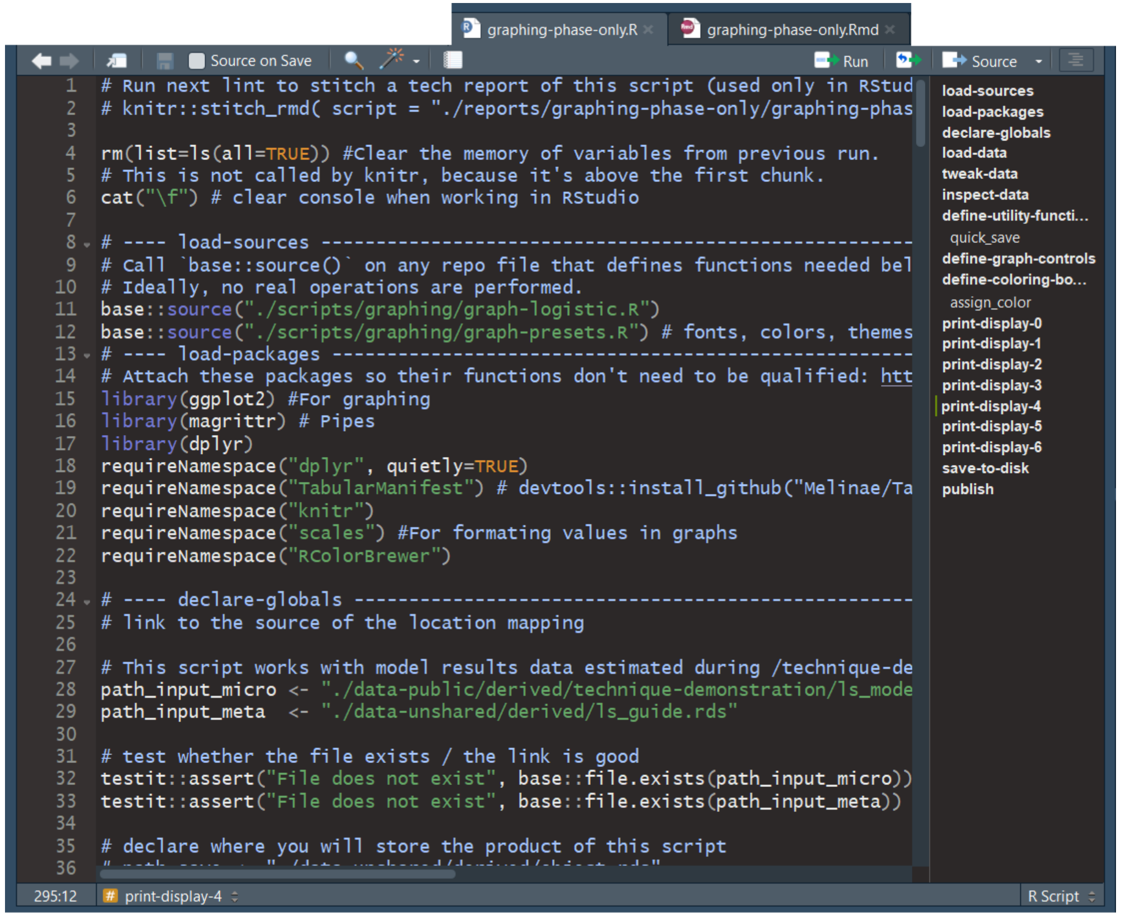
.Rmd
stores presentation
(how you tell about it)

.R + .Rmd = .html

B. Workflow Highlights

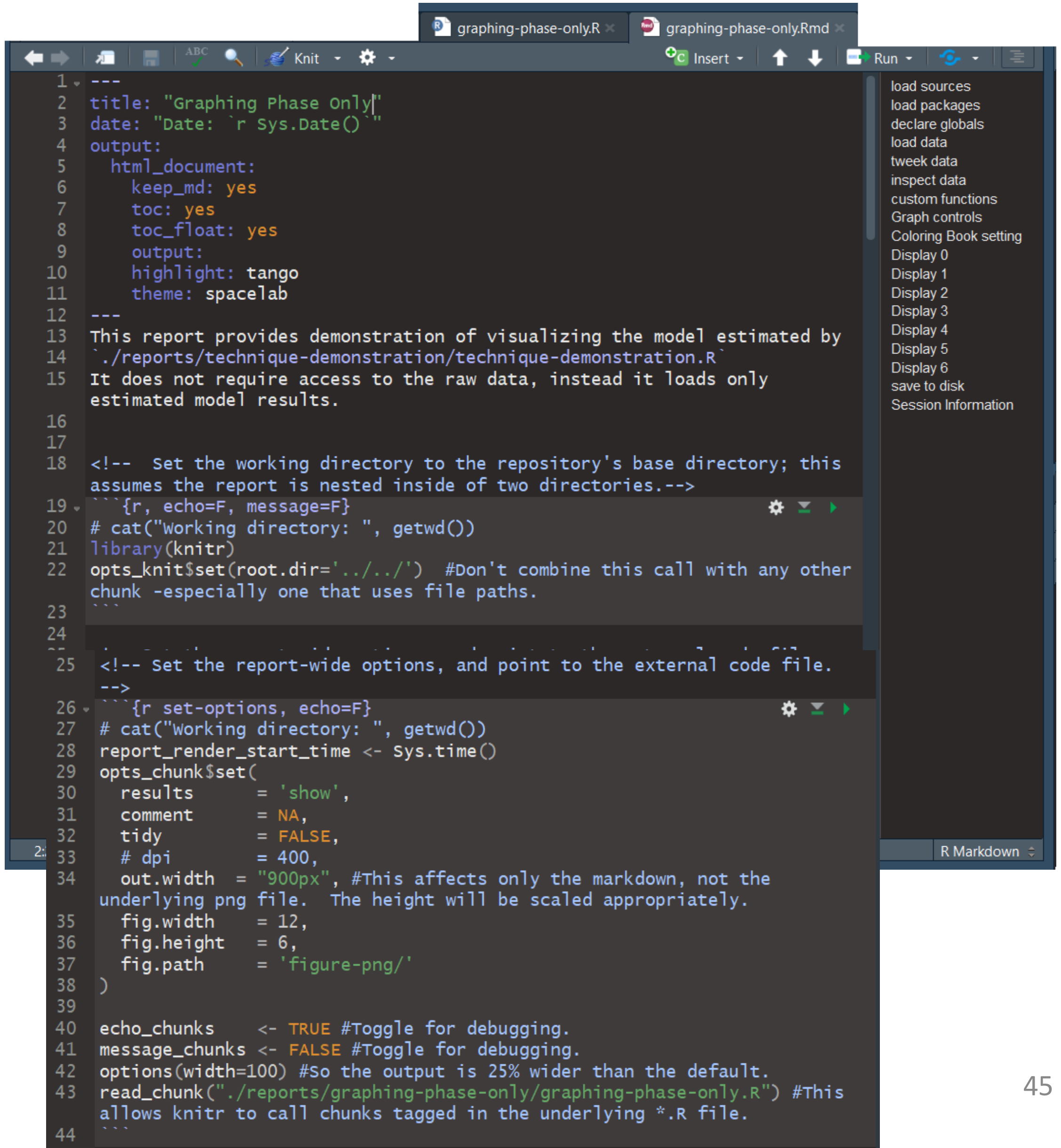
1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

.R – stores analysis (what really happens)



```r
# Run next lint to stitch a tech report of this script (used only in RStud
# knitr::stitch_rmd( script = "./reports/graphing-phase-only/graphing-phas

rm(list=ls(all=TRUE)) #Clear the memory of variables from previous run.
# This is not called by knitr, because it's above the first chunk.
cat("\f") # clear console when working in RStudio

# ---- load-sources ------------------------------------------------------
# Call `base::source()` on any repo file that defines functions needed bel
# Ideally, no real operations are performed.
base::source("./scripts/graphing/graph-logistic.R")
base::source("./scripts/graphing/graph-presets.R") # fonts, colors, themes
# ---- load-packages -----------------------------------------------------
# Attach these packages so their functions don't need to be qualified: htt
library(ggplot2) #For graphing
library(magrittr) # Pipes
library(dplyr)
requireNamespace("dplyr", quietly=TRUE)
requireNamespace("TabularManifest") # devtools::install_github("Melinae/Ta
requireNamespace("knitr")
requireNamespace("scales") #For formating values in graphs
requireNamespace("RColorBrewer")

# ---- declare-globals ---------------------------------------------------
# link to the source of the location mapping

# This script works with model results data estimated during /technique-de
path_input_micro <- "./data-public/derived/technique-demonstration/ls_mode
path_input_meta  <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
testit::assert("File does not exist", base::file.exists(path_input_micro))
testit::assert("File does not exist", base::file.exists(path_input_meta))

# declare where you will store the product of this script
# path_save <- "./data-unshared/derived/object.rds"
```

.Rmd – stores presentation (how you tell about it)



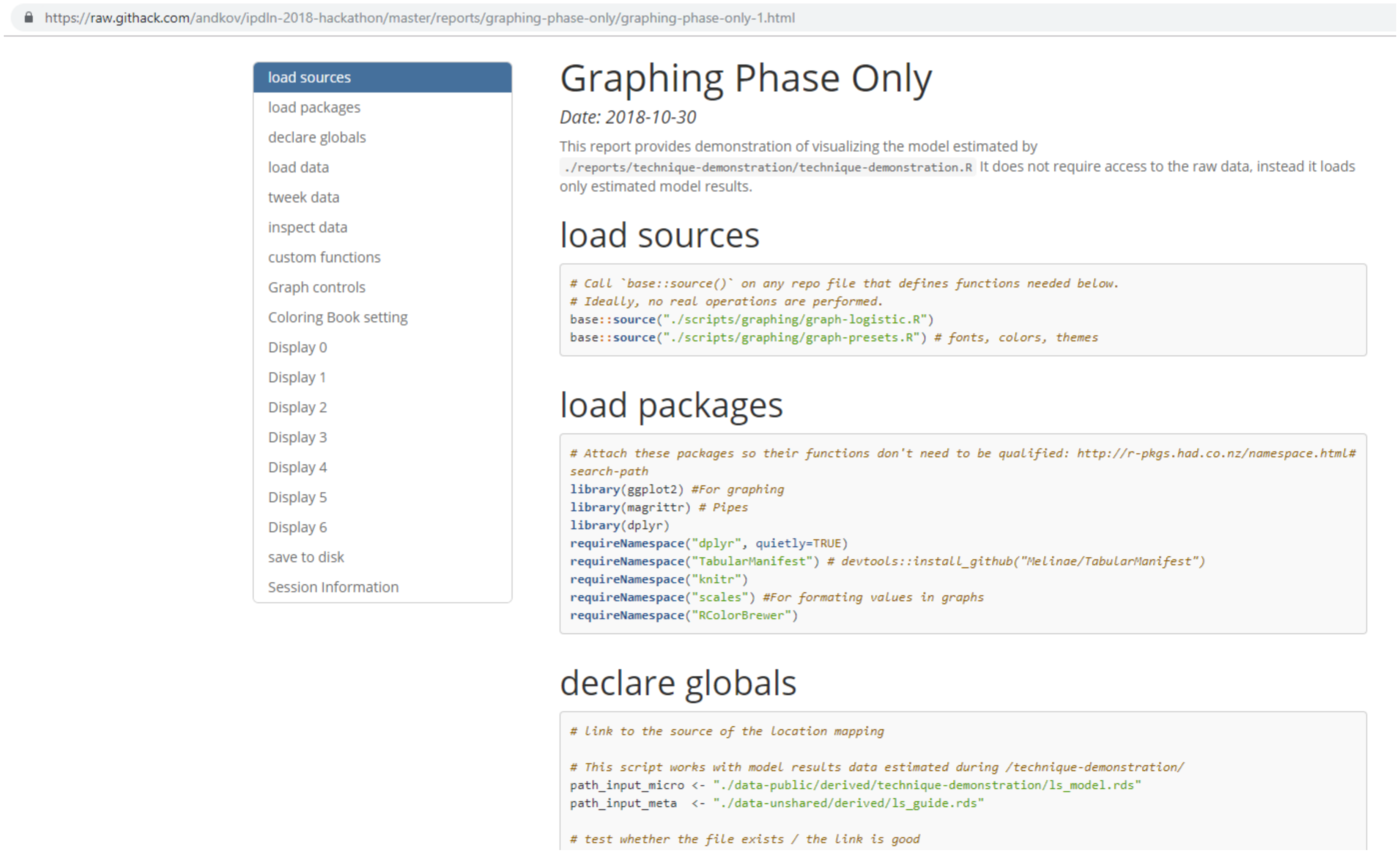
~~~
---
title: "Graphing Phase Only"
date: "Date: `r Sys.Date()`"
output:
  html_document:
    keep_md: yes
    toc: yes
    toc_float: yes
    output:
    highlight: tango
    theme: spacelab
---
This report provides demonstration of visualizing the model estimated by
`./reports/technique-demonstration/technique-demonstration.R`
It does not require access to the raw data, instead it loads only
estimated model results.


<!--  Set the working directory to the repository's base directory; this
assumes the report is nested inside of two directories.-->
```{r, echo=F, message=F}
# cat("Working directory: ", getwd())
library(knitr)
opts_knit$set(root.dir='../../')  #Don't combine this call with any other
chunk -especially one that uses file paths.
```

<!-- Set the report-wide options, and point to the external code file.
-->
```{r set-options, echo=F}
# cat("Working directory: ", getwd())
report_render_start_time <- Sys.time()
opts_chunk$set(
  results      = 'show',
  comment      = NA,
  tidy         = FALSE,
  # dpi        = 400,
  out.width  = "900px", #This affects only the markdown, not the
underlying png file.  The height will be scaled appropriately.
  fig.width    = 12,
  fig.height   = 6,
  fig.path     = 'figure-png/'
)

echo_chunks    <- TRUE #Toggle for debugging.
message_chunks <- FALSE #Toggle for debugging.
options(width=100) #So the output is 25% wider than the default.
read_chunk("./reports/graphing-phase-only/graphing-phase-only.R") #This
allows knitr to call chunks tagged in the underlying *.R file.
```
~~~

B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

.R – stores analysis (what really happens)

.Rmd – stores presentation (how you tell about it)



Graphing Phase Only

Date: 2018-10-30

This report provides demonstration of visualizing the model estimated by `./reports/technique-demonstration/technique-demonstration.R` It does not require access to the raw data, instead it loads only estimated model results.

load sources

```
# Call `base::source()` on any repo file that defines functions needed below.
# Ideally, no real operations are performed.
base::source("./scripts/graphing/graph-logistic.R")
base::source("./scripts/graphing/graph-presets.R") # fonts, colors, themes
```

load packages

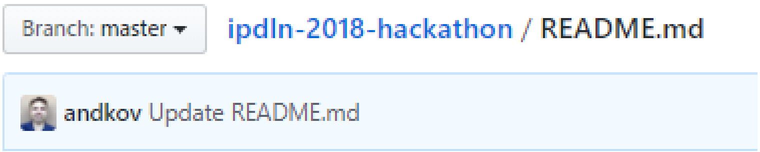
```
# Attach these packages so their functions don't need to be qualified: http://r-pkgs.had.co.nz/namespace.html#
search-path
library(ggplot2) #For graphing
library(magrittr) # Pipes
library(dplyr)
requireNamespace("dplyr", quietly=TRUE)
requireNamespace("TabularManifest") # devtools::install_github("Melinae/TabularManifest")
requireNamespace("knitr")
requireNamespace("scales") #For formating values in graphs
requireNamespace("RColorBrewer")
```

declare globals

```
# link to the source of the location mapping

# This script works with model results data estimated during /technique-demonstration/
path_input_micro <- "./data-public/derived/technique-demonstration/ls_model.rds"
path_input_meta  <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
```

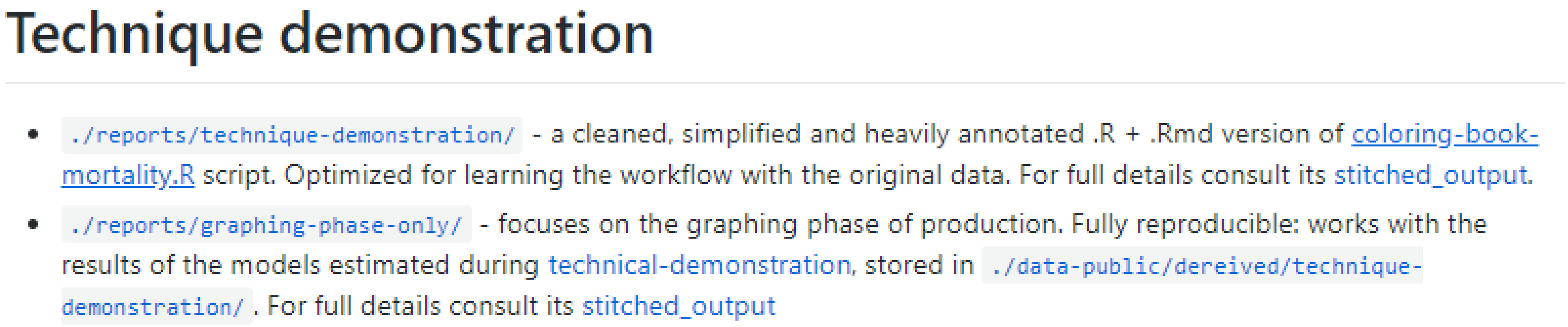
B. Workflow Highlights

1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html



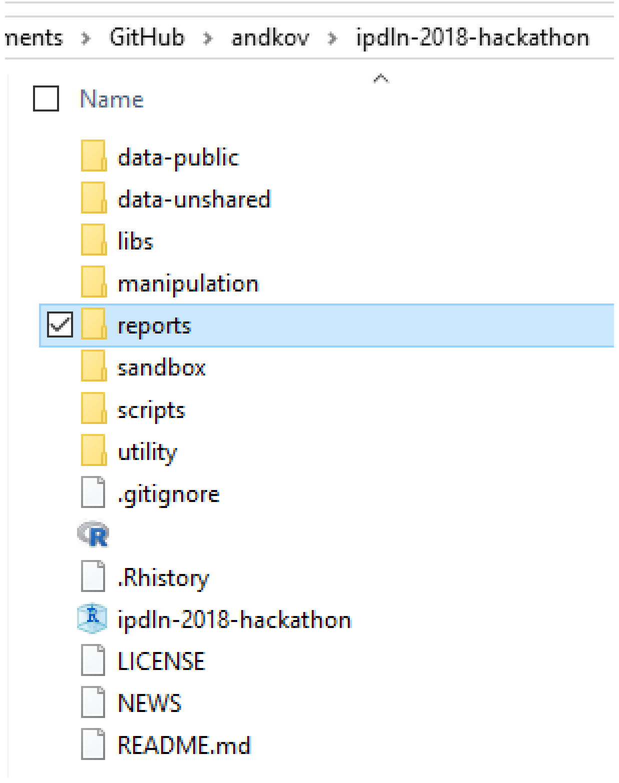
Branch: master ▾ ipdln-2018-hackathon / README.md

andkov Update README.md



Technique demonstration

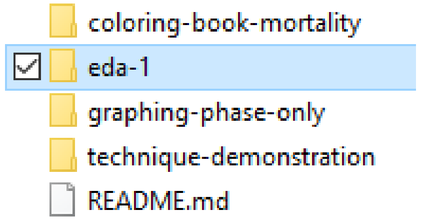
- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of coloring-book-mortality.R script. Optimized for learning the workflow with the original data. For full details consult its stitched_output.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during technical-demonstration, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its stitched_output



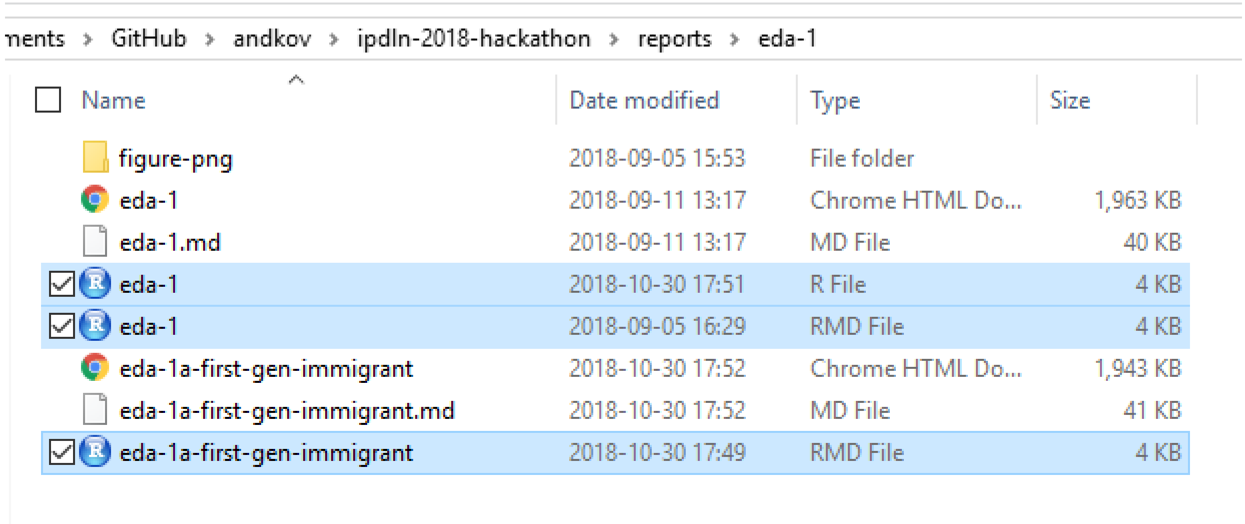
nents › GitHub › andkov › ipdln-2018-hackathon

□ Name

- data-public
- data-unshared
- libs
- manipulation
- ☑ reports
- sandbox
- scripts
- utility
- .gitignore
- .Rhistory
- ipdln-2018-hackathon
- LICENSE
- NEWS
- README.md



- coloring-book-mortality
- ☑ eda-1
- graphing-phase-only
- technique-demonstration
- README.md



nents › GitHub › andkov › ipdln-2018-hackathon › reports › eda-1

□ Name	Date modified	Type	Size
figure-png	2018-09-05 15:53	File folder	
eda-1	2018-09-11 13:17	Chrome HTML Do...	1,963 KB
eda-1.md	2018-09-11 13:17	MD File	40 KB
☑ eda-1	2018-10-30 17:51	R File	4 KB
☑ eda-1	2018-09-05 16:29	RMD File	4 KB
eda-1a-first-gen-immigrant	2018-10-30 17:52	Chrome HTML Do...	1,943 KB
eda-1a-first-gen-immigrant.md	2018-10-30 17:52	MD File	41 KB
☑ eda-1a-first-gen-immigrant	2018-10-30 17:49	RMD File	4 KB

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Autonomous phases: data cleaning, statistical modelling, graph production
1.3 Layers of Isolation: analysis vs presentation using .R + .Rmd = .html

Closing thoughts

- What makes “data science” a science? Reproducibility

- Principles to keep in mind
  - Scripts are better than GUIs
  - Notebooks are better than scripts
  - Projects are better than Notebooks

- “*There are only two hard things in programming: cache validation and naming things*” – Unknown
  - Success in Data Science = Craft + Imagination

Questions? Comments?

Andriy Koval

https://github.com/andkov

http://andriy.rbind.io