The Future of Data Science: Transparency and Equity

Dr. Shonn Cheng

2025-04-15

Who Am I

- Assistant Professor (Graduate Institute of Technological & Vocational Education at National Taipei University of Technology)
- Head of Curriculum Planning Committee in Teacher Education Center at National Taipei University of Technology
- Director of the META Lab

My Journey with Data Science

- *Ph.D. in Educational Studies from The Ohio State University (USA)*
- *M.A. in Quantitative Research, Evaluation and Measurement from The Ohio State University (USA)*
- M.A. in Curriculum & Instruction from New Mexico State University (USA)
- B.A. in English from Wenzao Ursuline University of Languages (Taiwan)
- A.M.S. in Business Administration from the National Taipei University of Business (Taiwan)

- Virginia Commonwealth University (08/2019-08/2020)
  - Training implementation and evaluation
  - Quantitative methodologist
  - Data analysis
  - Grant
  - Manuscripts

- Sam Houston State University (08/2020-2022/12)
  - Instructional Systems Design and Technology

  - Department of Library Science and Technology, College of Education
    * *Statistical Methods*
    * *Program Evaluation*

- National Taipei University of Technology (02/2023-present)
  - Graduate Institute of Technological and Vocational Education
  - College of Humanities & Social Sciences
    * *Statistis in Education*
    * *Training Implementation and Evaluation*

The Statistical Package War

Challenge 1: Transparency

- Transparency refers to the evaluation processes and conclusions being able to be scrutinised.
- They Studied Dishonesty. Was Their Work a Lie?
- Cornell Food Researcher's Downfall Raises Larger Questions For Science
- Amid a replication crisis in social science research, six-year study validates open science methods

Challenge 2: Equity

Equity and equality are often confused, but they represent distinct approaches to fairness. Equality means treating everyone the same, while equity means providing resources and opportunities tailored to individual needs to achieve equal outcomes. In essence, equality focuses on sameness, while equity focuses on fairness.

- Equity in online learning

Challenge 3: Data Science Flow

Discussion: What comes to your mind when it comes to data science?

Discussion: What is your normal procedure of doing data science?

Data science is "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" Provost & Fawcett (2013, p. 52).

R as a Promising Solution

- Open-Source and Free: Accessible to everyone without cost.
- Comprehensive Statistical Analysis: Supports advanced statistical methods and models.
- Extensive Libraries and Packages: Thousands of packages for specialized tasks.
- Data Manipulation and Cleaning: Powerful tools for transforming and cleaning data.
- Reproducibility and Transparency: Enables reproducible and transparent analyses with R Markdown.
- Advanced Data Visualization: High-quality, customizable plots with ggplot2.
- Cross-Platform: Works on Windows, macOS, and Linux.
- Integration with Other Tools: Integrates with Python, SQL, and big data tools.
- Active Community and Support: Large, helpful community with extensive resources.
- Suitable for Various Research Fields: Used widely in academia and diverse industries.
- Support for Reproducible Research: Encourages reproducibility through tools like R Markdown.
- Comprehensive Documentation: Detailed documentation for functions and packages.

Scenario

You are tasked with analyzing the relations between demographics, adverse childhood experiences (ACEs), and youth mental health outcomes. The dataset you are using comes from the 2017–2018 National Survey of Children's Health (NSCH), which is a nationally representative sample of children aged 0–17 years in the United States. You are expected to analyze the predictive relationss between demographic variables (age, sex, race, household income), ACEs, and parent-reported mental health conditions (depression, anxiety) and behavioral problems.

import

```
#install.packages("tidyverse")
#install.packages("haven")
#install.packages("psych")
#install.packages("fastDummies")
#install.packages("survey")
library(tidyverse)
```

```r
library(haven)
library(psych)
library(fastDummies)
library(survey)
```

```
Warning: package 'survival' was built under R version 4.3.3
```

```r
data<-read_dta("data.dta")
```

tidy

```r
head(data)
```

```
# A tibble: 6 x 753
  HHID      FIPSST    STRATUM FORMTYPE TOTKIDS_R HHLANGUAGE SC_AGE_YEARS SC_SEX
  <dbl+lbl> <dbl+lbl>   <dbl>    <dbl> <dbl+lbl> <dbl+lbl>  <dbl+lbl>    <dbl+l>
1 17000010  37  [Nort~       1        1 3  [3]      3  [Other]   0           2 [Fem~
2 17000013   2  [Alas~       2        3 1  [1]      1  [Englis~ 13           2 [Fem~
3 17000025  40  [Okla~       1        3 1  [1]      1  [Englis~ 15           1 [Mal~
4 17000031  13  [Geor~       1        2 1  [1]      1  [Englis~  9           1 [Mal~
5 17000034  31  [Nebr~       1        2 2  [2]      1  [Englis~  8           2 [Fem~
6 17000044  13  [Geor~       1        1 2  [2]      1  [Englis~  4           1 [Mal~
# i 745 more variables: K2Q35A_1_YEARS <dbl+lbl>, MOMAGE <dbl+lbl>,
#   K6Q41R_STILL <dbl+lbl>, K6Q42R_NEVER <dbl+lbl>, K6Q43R_NEVER <dbl+lbl>,
#   K6Q13A <dbl+lbl>, K6Q13B <dbl+lbl>, K6Q14A <dbl+lbl>, K6Q14B <dbl+lbl>,
#   K4Q32X01 <dbl+lbl>, K4Q32X02 <dbl+lbl>, K4Q32X03 <dbl+lbl>,
#   K4Q32X04 <dbl+lbl>, K4Q32X05 <dbl+lbl>, DENTALSERV1 <dbl+lbl>,
#   DENTALSERV2 <dbl+lbl>, DENTALSERV3 <dbl+lbl>, DENTALSERV4 <dbl+lbl>,
#   DENTALSERV5 <dbl+lbl>, DENTALSERV6 <dbl+lbl>, DENTALSERV7 <dbl+lbl>, ...
```

transform: age

```r
table(data$SC_AGE_YEARS)
```

```

   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15 
1778 2148 2974 2671 2563 2603 2392 2430 2565 2677 2869 2963 2920 3204 3476 3639 
  16   17 
4117 4140 
```

transform: sex

```
#1 "male"
#2 "female"
table(data$SC_SEX)
```

```
    1     2 
27044 25085 
```

transform: race

```
#1 "Hispanic"
#2 "White, non-Hispanic"
#3 "Black, non-Hispanic"
#4 "Asian, non-Hispanic"
#5 "American Indian or Alaskan Native, non-Hispanic"
#6 "Native Hawaiian and Other Pacific Islander, non-Hispanic"
#7 "Multi Race, Non-Hispanic"
table(data$SC_RACE_R)
```

```
    1     2     3     4     5     6     7 
39947  3527   414  2623   140  1407  4071 
```

```
data <- dummy_cols(data, select_columns = "SC_RACE_R",
                   remove_first_dummy = FALSE)
```

transform: household income

```
#1 "0-99% FPL"
#2 "100%-199% FPL"
#3 "200%-399% FPL"
#4 "400% FPL or above"
table(data$povlev4_1718)
```

```
   1     2     3     4 
 6355  8270 15883 21621 
```

transform: adverse childhood experiences

```
table(data$ACEct_1718)
```

```
    0     1     2     3     4     5     6     7     8     9    99 
31159 10957  4355  2226  1261   732   437   196    55    11   740 
```

```
data<-data %>% mutate(ACEct_1718_r = case_when(
  ACEct_1718 == 99 ~ NA_real_,  # Replace 99 with NA
  TRUE ~ as.numeric(ACEct_1718)  # Otherwise, convert to numeric
))
```

transform: depression

```
#1 = Yes; 2 = No;
table(data$K2Q32A)
```

```
    1     2    99 
 2550 49395   184 
```

```
data<-data %>% mutate(K2Q32A_r = case_when(
  K2Q32A == 99 ~ NA_real_,  # Replace 99 with NA
  K2Q32A == 2 ~ 0,
  K2Q32A == 1 ~ 1
))
```

transform: anxiety

```r
table(data$K2Q33A)
```

```

   1     2    99
 5289 46670   170
```

```r
data<-data %>% mutate(K2Q33A_r = case_when(
  K2Q33A == 99 ~ NA_real_,  # Replace 99 with NA
  K2Q33A == 2 ~ 0,
  K2Q33A == 1 ~ 1
))
```

transform: behavioral problems

```r
table(data$K2Q34A)
```

```

   1     2    99
 4265 47710   154
```

```r
data<-data %>% mutate(K2Q34A_r = case_when(
  K2Q34A == 99 ~ NA_real_,  # Replace 99 with NA
  K2Q34A == 2 ~ 0,
  K2Q34A == 1 ~ 1
))
```

transform: create unique strata

```r
data <- data %>%
  mutate(
    FIPSST = as_factor(FIPSST),  # Convert to factor if necessary
    STRATUM = as_factor(STRATUM),  # Convert to factor if necessary
  ) %>%
  group_by(FIPSST, STRATUM) %>%  # Group by the two variables
  mutate(stratacross = cur_group_id()) %>%  # Create unique strata
  ungroup()  # Ungroup after the mutation
```

```r
# Check the result
table(data$stratacross)
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16 
 980   84  794  158  943  106  980  132  933   77  935   79  944   63  940   58 
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32 
 879   82  925   93  942   86  639  335  978  102  913   83  925   82  957   59 
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48 
1019   72  986   79  966   92  957   58  922   73  970   53  944   66 1011   39 
  49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64 
 996  114  995   65  893   84  950   49  934  109  993   65  933   66  876  143 
  65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80 
 870  108  966   80  973   72  952   69  972  121  896   66  950   53  946   79 
  81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96 
 983   67  925   75  920   93  917   84  934   68  918   98  940   60  896   85 
  97   98   99  100  101  102 
 935  107  971   58  858  106 
```

transform: subset

```r
df<-data %>%
  select(HHID, stratacross, FWC_1718,
         SC_AGE_YEARS, SC_SEX, SC_RACE_R,
         SC_RACE_R_1:SC_RACE_R_7, povlev4_1718,
         ACEct_1718_r, K2Q32A_r, K2Q33A_r, K2Q34A_r)

df %>% describe()
```

```
             vars     n        mean        sd     median    trimmed       mad
HHID            1 52129 17672095.48 496610.11 18025136.0 17692447.12 185723.82
stratacross     2 52129       50.94     29.35       51.0       50.92     38.55
FWC_1718        3 52129     1408.60   2739.23      638.6      870.03    688.48
SC_AGE_YEARS    4 52129        9.45      5.24       10.0        9.60      7.41
SC_SEX          5 52129        1.48      0.50        1.0        1.48      0.00
SC_RACE_R       6 52129        1.85      1.83        1.0        1.34      0.00
SC_RACE_R_1     7 52129        0.77      0.42        1.0        0.83      0.00
SC_RACE_R_2     8 52129        0.07      0.25        0.0        0.00      0.00
SC_RACE_R_3     9 52129        0.01      0.09        0.0        0.00      0.00
```

```
SC_RACE_R_4     10 52129        0.05       0.22        0.0         0.00       0.00
SC_RACE_R_5     11 52129        0.00       0.05        0.0         0.00       0.00
SC_RACE_R_6     12 52129        0.03       0.16        0.0         0.00       0.00
SC_RACE_R_7     13 52129        0.08       0.27        0.0         0.00       0.00
povlev4_1718    14 52129        3.01       1.03        3.0         3.14       1.48
ACEct_1718_r    15 51389        0.77       1.29        0.0         0.47       0.00
K2Q32A_r        16 51945        0.05       0.22        0.0         0.00       0.00
K2Q33A_r        17 51959        0.10       0.30        0.0         0.00       0.00
K2Q34A_r        18 51975        0.08       0.27        0.0         0.00       0.00
                      min          max   range  skew kurtosis      se
HHID          17000010.00 18176036.00 1176026 -0.34    -1.84 2175.08
stratacross          1.00      102.00     101  0.00    -1.19    0.13
FWC_1718             9.34    56123.34   56114  7.64    93.77   12.00
SC_AGE_YEARS         0.00       17.00      17 -0.19    -1.23    0.02
SC_SEX               1.00        2.00       1  0.08    -1.99    0.00
SC_RACE_R            1.00        7.00       6  2.07     2.73    0.01
SC_RACE_R_1          0.00        1.00       1 -1.26    -0.42    0.00
SC_RACE_R_2          0.00        1.00       1  3.44     9.85    0.00
SC_RACE_R_3          0.00        1.00       1 11.09   120.92    0.00
SC_RACE_R_4          0.00        1.00       1  4.11    14.93    0.00
SC_RACE_R_5          0.00        1.00       1 19.22   367.34    0.00
SC_RACE_R_6          0.00        1.00       1  5.84    32.08    0.00
SC_RACE_R_7          0.00        1.00       1  3.14     7.89    0.00
povlev4_1718         1.00        4.00       3 -0.69    -0.73    0.00
ACEct_1718_r         0.00        9.00       9  2.27     5.80    0.01
K2Q32A_r             0.00        1.00       1  4.17    15.42    0.00
K2Q33A_r             0.00        1.00       1  2.63     4.94    0.00
K2Q34A_r             0.00        1.00       1  3.05     7.28    0.00
```

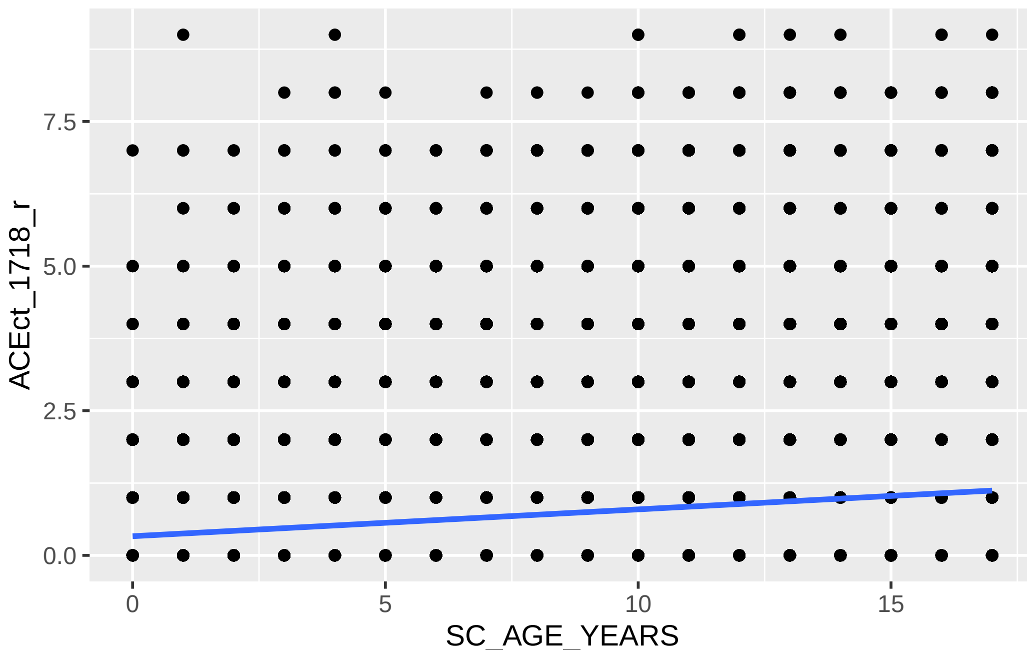
transform: create suvery object

```
sd<-svydesign(id=~HHID, strata=~stratacross, weights=~FWC_1718, data=df)
```

visualize

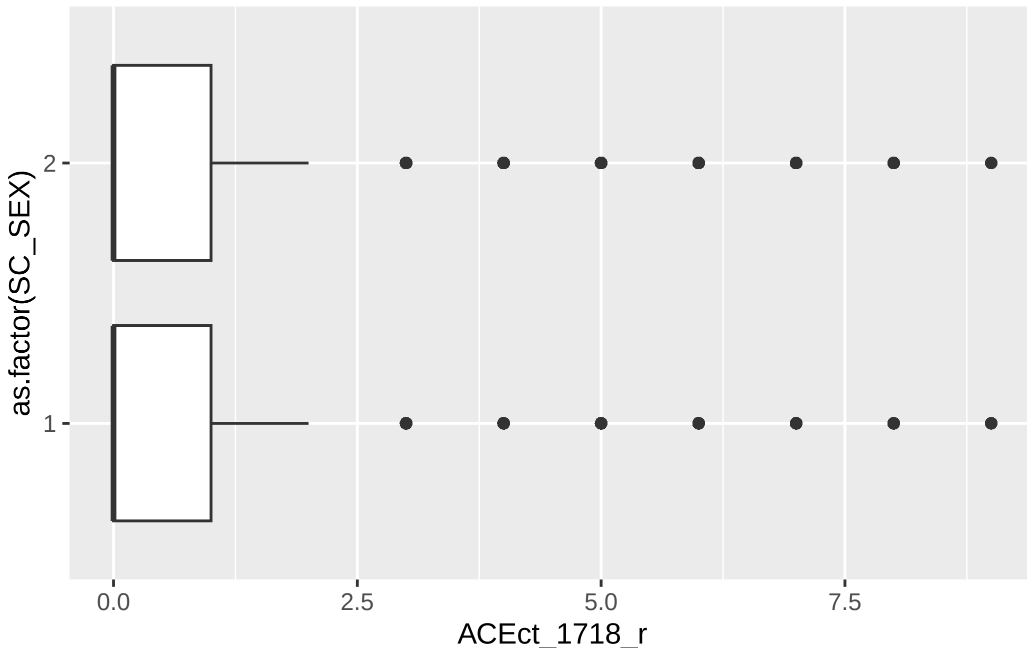
```
df %>%
  ggplot(mapping = aes(x = SC_AGE_YEARS, y = ACEct_1718_r)) +
  geom_point() +  # scatter plot
  geom_smooth(method = "lm", se = FALSE)  # linear regression line
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 740 rows containing missing values or values outside the scale range
(`geom_point()`).
```



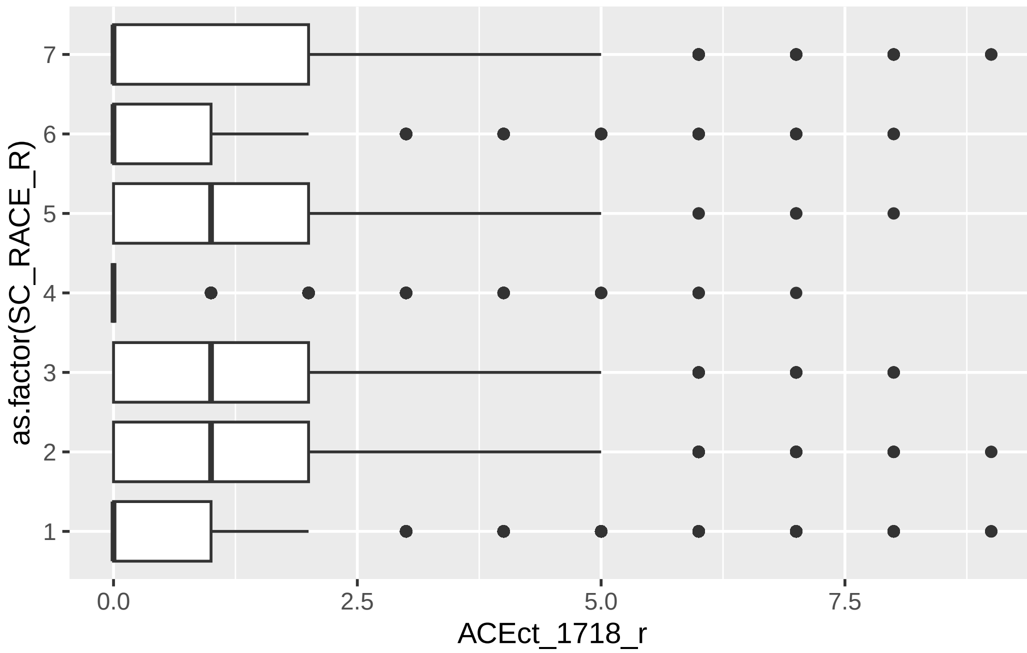

```r
df %>%
  ggplot(mapping=aes(x = ACEct_1718_r, y = as.factor(SC_SEX))) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



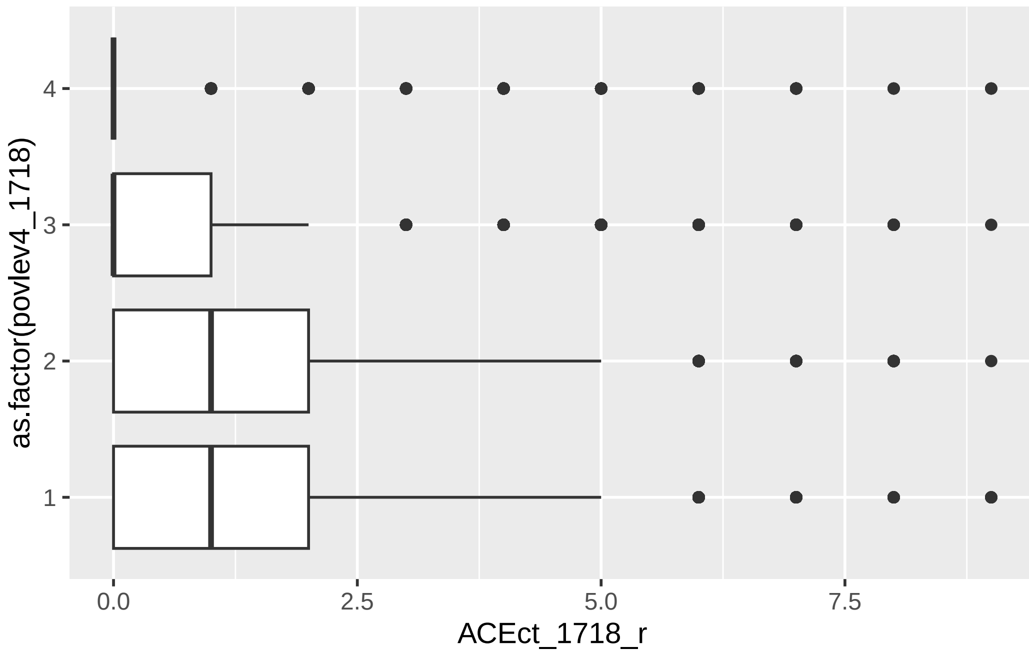

```
df %>%
  ggplot(mapping=aes(x = ACEct_1718_r, y = as.factor(SC_RACE_R))) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



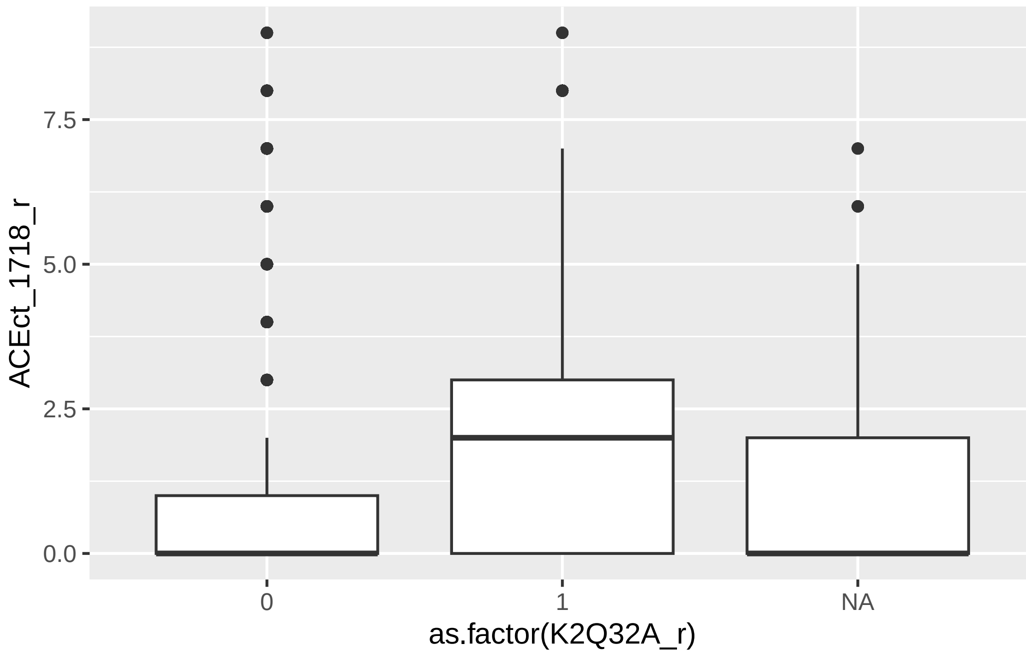

```
df %>%
  ggplot(mapping=aes(x = ACEct_1718_r, y = as.factor(povlev4_1718))) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



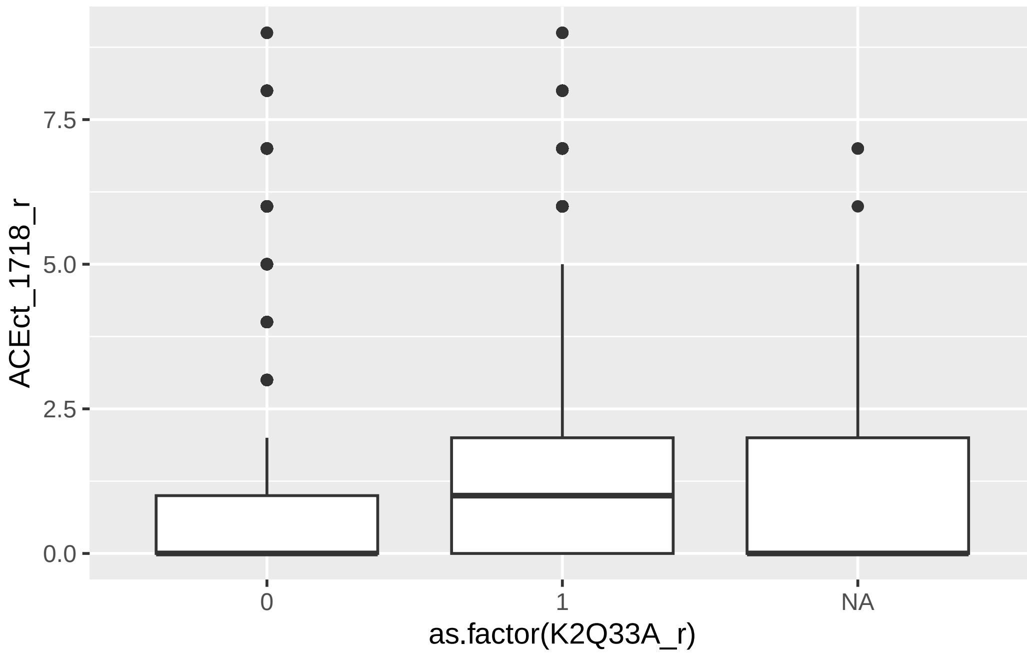

```
df %>%
  ggplot(mapping=aes(x = as.factor(K2Q32A_r), ACEct_1718_r)) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



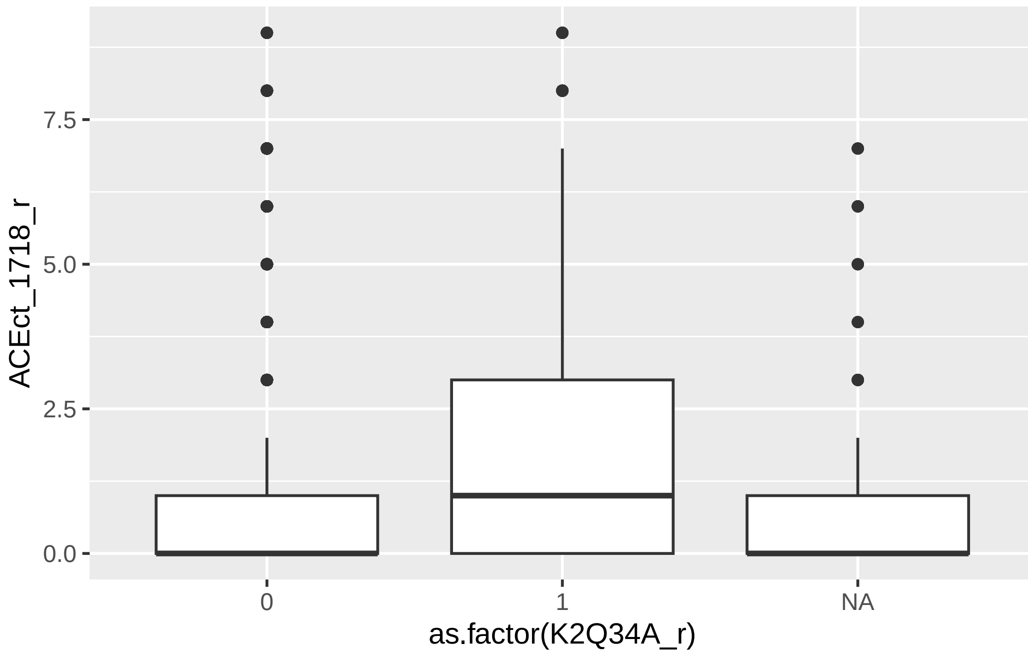

```
df %>%
  ggplot(mapping=aes(x = as.factor(K2Q33A_r), ACEct_1718_r)) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




```
df %>%
  ggplot(mapping=aes(x = as.factor(K2Q34A_r), ACEct_1718_r)) +
  geom_boxplot()
```

```
Warning: Removed 740 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



model: predict adverse childhood experiences

```
#multiple linear regression

# Fit a linear regression model (Gaussian family)
fit <- svyglm(ACEct_1718_r ~ SC_AGE_YEARS + as.factor(SC_SEX) +
              SC_RACE_R_1 + SC_RACE_R_2 +
              SC_RACE_R_4 + SC_RACE_R_5 +
              SC_RACE_R_6 + SC_RACE_R_7 +
              as.factor(povlev4_1718),
              data = df, design = sd, family = gaussian())

# Show the summary of the model
summary(fit)
```

```
Call:
svyglm(formula = ACEct_1718_r ~ SC_AGE_YEARS + as.factor(SC_SEX) +
    SC_RACE_R_1 + SC_RACE_R_2 + SC_RACE_R_4 + SC_RACE_R_5 + SC_RACE_R_6 +
    SC_RACE_R_7 + as.factor(povlev4_1718), design = sd, family = gaussian(),
    data = df)
```

```
Survey design:
svydesign(id = ~HHID, strata = ~stratacross, weights = ~FWC_1718,
    data = df)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.959366   0.151854   6.318 2.68e-10 ***
SC_AGE_YEARS              0.050242   0.002047  24.539  < 2e-16 ***
as.factor(SC_SEX)2       -0.005071   0.023268  -0.218 0.827484    
SC_RACE_R_1              -0.202997   0.146451  -1.386 0.165720    
SC_RACE_R_2              -0.006130   0.151014  -0.041 0.967622    
SC_RACE_R_4              -0.536260   0.149321  -3.591 0.000329 ***
SC_RACE_R_5              -0.021917   0.237976  -0.092 0.926620    
SC_RACE_R_6              -0.359457   0.163355  -2.200 0.027779 *  
SC_RACE_R_7               0.190450   0.154941   1.229 0.219012    
as.factor(povlev4_1718)2 -0.158850   0.048153  -3.299 0.000971 ***
as.factor(povlev4_1718)3 -0.466940   0.040726 -11.465  < 2e-16 ***
as.factor(povlev4_1718)4 -0.811163   0.038993 -20.803  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.497997)

Number of Fisher Scoring iterations: 2
```

model: predict mental health and behavior programs

```
#logistic regression
fit <- svyglm(K2Q32A_r ~ ACEct_1718_r,
              data = df, design = sd, family = quasibinomial())
summary(fit)
```

```
Call:
svyglm(formula = K2Q32A_r ~ ACEct_1718_r, design = sd, family = quasibinomial(),
    data = df)

Survey design:
svydesign(id = ~HHID, strata = ~stratacross, weights = ~FWC_1718,
```

```
    data = df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.99989    0.06170  -64.83   <2e-16 ***
ACEct_1718_r  0.51891    0.02098   24.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.9250922)

Number of Fisher Scoring iterations: 6
```

```r
fit <- svyglm(K2Q33A_r ~ ACEct_1718_r,
              data = df, design = sd, family = quasibinomial())
summary(fit)
```

```
Call:
svyglm(formula = K2Q33A_r ~ ACEct_1718_r, design = sd, family = quasibinomial(),
    data = df)

Survey design:
svydesign(id = ~HHID, strata = ~stratacross, weights = ~FWC_1718,
    data = df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.88348    0.03864  -74.63   <2e-16 ***
ACEct_1718_r  0.36376    0.01668   21.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.9732264)

Number of Fisher Scoring iterations: 5
```

```r
fit <- svyglm(K2Q34A_r ~ ACEct_1718_r,
              data = df, design = sd, family = quasibinomial())
summary(fit)
```

```
Call:
svyglm(formula = K2Q34A_r ~ ACEct_1718_r, design = sd, family = quasibinomial(),
    data = df)

Survey design:
svydesign(id = ~HHID, strata = ~stratacross, weights = ~FWC_1718,
    data = df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.06156    0.04568  -67.02   <2e-16 ***
ACEct_1718_r  0.46950    0.01888   24.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.9647003)

Number of Fisher Scoring iterations: 5
```

communicate

The analysis reveals several key insights into the predictors of adverse childhood experiences (ACEs). First, older children tend to have higher ACE scores, suggesting that the accumulation of adverse experiences increases with age. Second, racial and ethnic disparities in ACEs are evident, with Asian children and Native Hawaiian or Other Pacific Islander children experiencing significantly lower ACEs compared to their Black counterparts. Additionally, socioeconomic status plays a critical role, as children from wealthier families tend to report fewer ACEs, highlighting the protective effects of financial stability on childhood well-being. Finally, higher ACE scores are strongly associated with increased mental health challenges, including higher rates of depression, anxiety, and behavioral problems, emphasizing the long-term psychological impact of childhood adversity.

communicate: bonus

Shiny

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508