Statistics Project Markdown

B150297

2023-11-20

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
##
##
##
## Please cite as:
##
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
##
##
##
## Attaching package: 'boot'
##
##
## The following object is masked from 'package:car':
##
##     logit
```

```
## Warning: package 'margins' was built under R version 4.3.2
```

```
## [1] 72390  6693
```

```
## [1] 72390     6
```

```
## tibble [72,390 x 6] (S3: tbl_df/tbl/data.frame)
##  $ EDUC     : num [1:72390] 16 10 12 17 12 14 13 16 12 12 ...
##   ..- attr(*, "label")= chr "Highest year of school completed"
##   ..- attr(*, "format.sas")= chr "EDUC"
##  $ AGE      : num [1:72390] 23 70 48 27 61 26 28 27 21 30 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "format.sas")= chr "AGE"
##  $ VAXSAFE  : num [1:72390] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "VACCINES ARE SAFE"
##   ..- attr(*, "format.sas")= chr "AGREESCALE2F"
##  $ VAXKIDS  : num [1:72390] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "VACCINES ARE IMPORTANT FOR CHILDREN"
##   ..- attr(*, "format.sas")= chr "AGREESCALE2F"
##  $ COVID12  : num [1:72390] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "COVID VACCINE EVER"
##   ..- attr(*, "format.sas")= chr "YESNO"
##  $ VAXDOHARM: num [1:72390] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "VACCINES DO MORE HARM THAN GOOD"
##   ..- attr(*, "format.sas")= chr "AGREESCALE2C"
```

```
##       EDUC            AGE           VAXSAFE         VAXKIDS
##  Min.   : 0.00   Min.   :18.00   Min.   :1.00    Min.   :1.00
##  1st Qu.:12.00   1st Qu.:32.00   1st Qu.:1.00    1st Qu.:1.00
##  Median :12.00   Median :44.00   Median :2.00    Median :1.00
##  Mean   :13.03   Mean   :46.56   Mean   :2.06    Mean   :1.66
##  3rd Qu.:16.00   3rd Qu.:60.00   3rd Qu.:3.00    3rd Qu.:2.00
##  Max.   :20.00   Max.   :89.00   Max.   :5.00    Max.   :5.00
##  NA's   :263     NA's   :769     NA's   :71158   NA's   :71157
##     COVID12        VAXDOHARM
##  Min.   :1.00    Min.   :1.00
##  1st Qu.:1.00    1st Qu.:3.00
##  Median :1.00    Median :4.00
##  Mean   :1.19    Mean   :3.78
##  3rd Qu.:1.00    3rd Qu.:5.00
##  Max.   :2.00    Max.   :5.00
##  NA's   :71164   NA's   :71274
```

```
## data
##
##  6  Variables      72390  Observations
## --------------------------------------------------------------------------------
## EDUC : Highest year of school completed  Format:EDUC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    72127      263       21    0.969    13.03     3.45        8        9
##      .25      .50      .75      .90      .95
##       12       12       16       17       18
##
## lowest :  0  1  2  3  4, highest: 16 17 18 19 20
## --------------------------------------------------------------------------------
## AGE : Age of respondent  Format:AGE
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    71621      769       72        1    46.56    20.11       22       25
##      .25      .50      .75      .90      .95
##       32       44       60       72       78
##
## lowest : 18 19 20 21 22, highest: 85 86 87 88 89
## --------------------------------------------------------------------------------
## VAXSAFE : VACCINES ARE SAFE  Format:AGREESCALE2F
##        n  missing distinct     Info     Mean      Gmd
##     1232    71158        5    0.902    2.058    1.008
##
## Value          1     2     3     4     5
## Frequency    407   430   332    43    20
## Proportion 0.330 0.349 0.269 0.035 0.016
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## VAXKIDS : VACCINES ARE IMPORTANT FOR CHILDREN  Format:AGREESCALE2F
##        n  missing distinct     Info     Mean      Gmd
##     1233    71157        5    0.813    1.659   0.8403
##
## Value          1     2     3     4     5
## Frequency    661   388   143    26    15
## Proportion 0.536 0.315 0.116 0.021 0.012
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## COVID12 : COVID VACCINE EVER  Format:YESNO
##        n  missing distinct     Info     Mean      Gmd
##     1226    71164        2    0.453    1.185    0.302
##
## Value          1     2
## Frequency    999   227
## Proportion 0.815 0.185
## --------------------------------------------------------------------------------
## VAXDOHARM : VACCINES DO MORE HARM THAN GOOD  Format:AGREESCALE2C
##        n  missing distinct     Info     Mean      Gmd
##     1116    71274        5    0.918     3.78    1.173
##
## Value          1     2     3     4     5
## Frequency     48    88   259   388   333
```

```
## Proportion 0.043 0.079 0.232 0.348 0.298
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------------------
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
##   177    49   158   268   326   410   866   896  2786  2172  3010  3942 21401
##    13    14    15    16    17    18    19    20
##  5905  8208  3307  9994  2392  2945  1112  1803
```

```
##
##   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
##  267  904  951 1084 1159 1321 1296 1469 1438 1472 1548 1451 1571 1464 1566 1526
##   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49
## 1552 1503 1492 1481 1466 1373 1412 1346 1345 1341 1289 1199 1236 1211 1193 1227
##   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65
## 1148 1191 1128 1145 1106 1054 1137 1046 1078 1047 1085  975 1002  968  854  952
##   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81
##  874  933  877  818  857  719  722  634  705  595  572  546  491  446  389  375
##   82   83   84   85   86   87   88   89
##  312  290  268  221  211  158  130  409
```

```
##
##   1   2   3   4   5
## 407 430 332  43  20
```

```
##
##   1   2   3   4   5
## 661 388 143  26  15
```

```
##
##   1   2
## 999 227
```

```
##
##   1   2   3   4   5
##  48  88 259 388 333
```

```
## [1] 285785
```

```
## [1] 71158
```

```
## [1] 98.29811
```

```
## [1] 98.29673
```

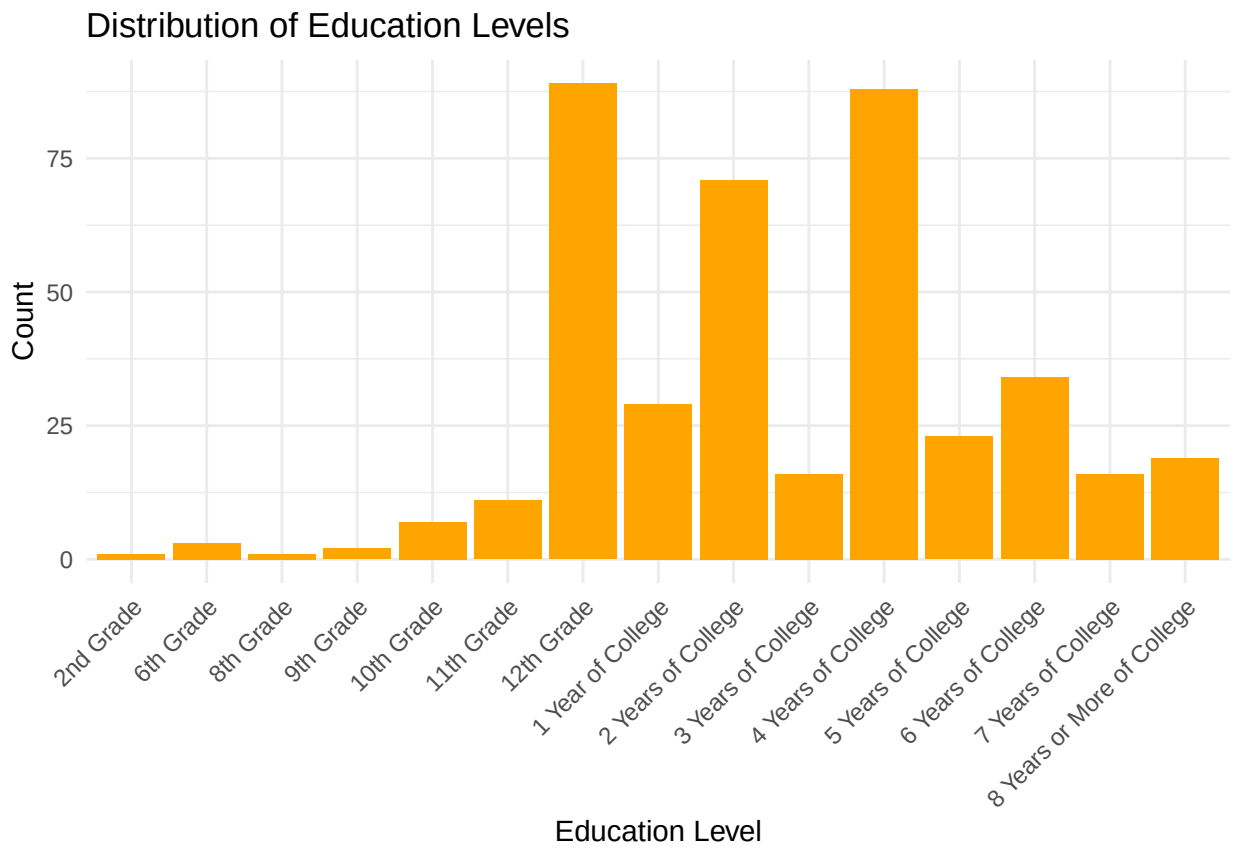
```
## [1] 98.3064
```

```
## [1] 98.45835
```

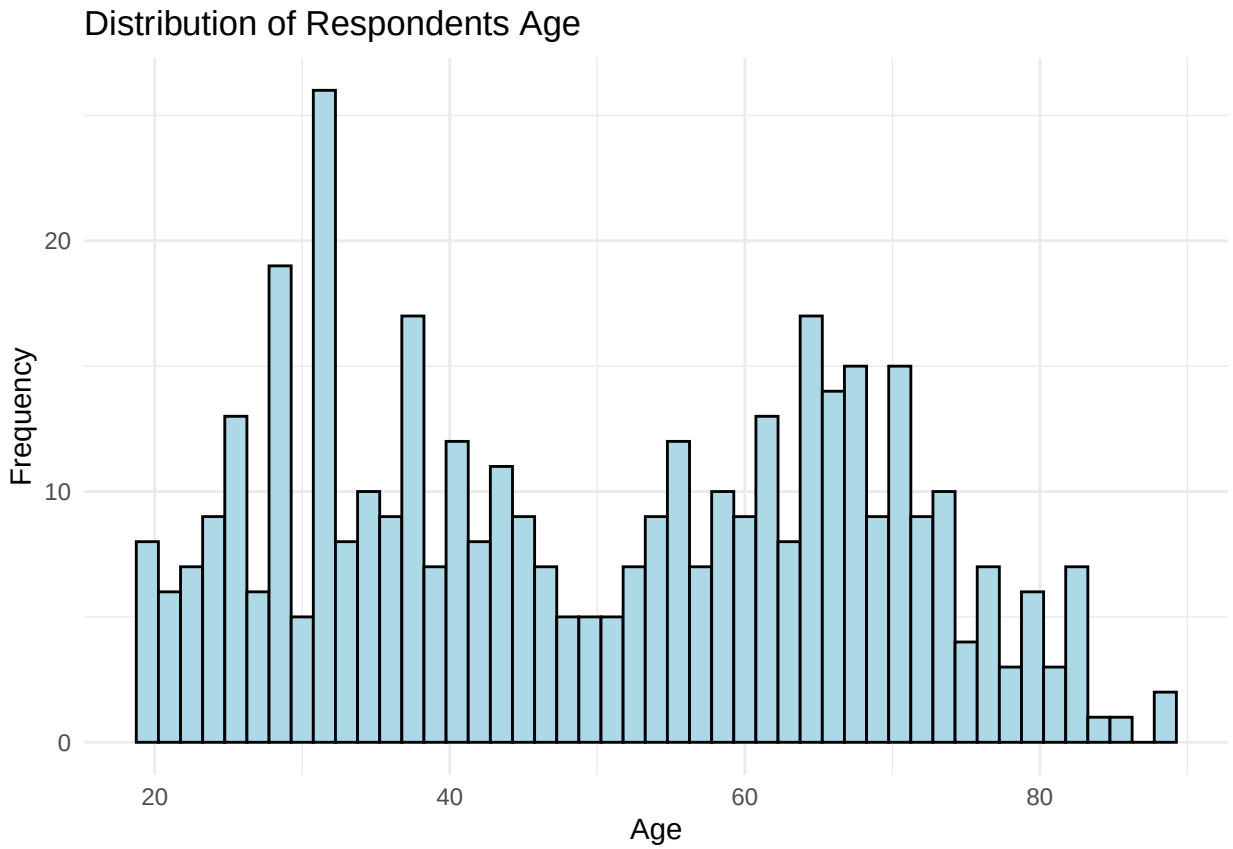
```
## [1] 0

## data_complete
##
##  6  Variables      410  Observations
## --------------------------------------------------------------------------------
## EDUC : Highest year of school completed  Format:EDUC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      410        0       15    0.973    14.66    3.024       11       12
##      .25      .50      .75      .90      .95
##       12       14       16       18       19
##
## Value          2     6     8     9    10    11    12    13    14    15    16
## Frequency      1     3     1     2     7    11    89    29    71    16    88
## Proportion 0.002 0.007 0.002 0.005 0.017 0.027 0.217 0.071 0.173 0.039 0.215
##
## Value         17    18    19    20
## Frequency     23    34    16    19
## Proportion 0.056 0.083 0.039 0.046
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## AGE : Age of respondent  Format:AGE
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      410        0       69        1    49.84    20.91    23.45    26.00
##      .25      .50      .75      .90      .95
##    33.00    50.00    66.00    73.00    78.00
##
## lowest : 19 20 21 22 23, highest: 83 84 86 88 89
## --------------------------------------------------------------------------------
## VAXSAFE : VACCINES ARE SAFE  Format:AGREESCALE2F
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5    0.899    2.015   0.9882
##
## Value          1     2     3     4     5
## Frequency    144   137   112    13     4
## Proportion 0.351 0.334 0.273 0.032 0.010
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## VAXKIDS : VACCINES ARE IMPORTANT FOR CHILDREN  Format:AGREESCALE2F
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5    0.823    1.673   0.8381
##
## Value          1     2     3     4     5
## Frequency    214   130    56     6     4
## Proportion 0.522 0.317 0.137 0.015 0.010
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## COVID12 : COVID VACCINE EVER  Format:YESNO
##        n  missing distinct     Info     Mean      Gmd
##      410        0        2    0.476    1.198   0.3178
##
```

```
## Value          1     2
## Frequency    329    81
## Proportion 0.802 0.198
## ---------------------------------------------------------------------------
## VAXDOHARM : VACCINES DO MORE HARM THAN GOOD  Format:AGREESCALE2C
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5      0.9     3.98    1.105
##
## Value          1     2     3     4     5
## Frequency     13    18    94   124   161
## Proportion 0.032 0.044 0.229 0.302 0.393
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------

## data
##
##  6  Variables      410  Observations
## ---------------------------------------------------------------------------
## EDUC : Highest year of school completed  Format:EDUC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      410        0       15    0.973    14.66    3.024       11       12
##      .25      .50      .75      .90      .95
##       12       14       16       18       19
##
## Value          2     6     8     9    10    11    12    13    14    15    16
## Frequency      1     3     1     2     7    11    89    29    71    16    88
## Proportion 0.002 0.007 0.002 0.005 0.017 0.027 0.217 0.071 0.173 0.039 0.215
##
## Value         17    18    19    20
## Frequency     23    34    16    19
## Proportion 0.056 0.083 0.039 0.046
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## AGE : Age of respondent  Format:AGE
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      410        0       69        1    49.84    20.91    23.45    26.00
##      .25      .50      .75      .90      .95
##    33.00    50.00    66.00    73.00    78.00
##
## lowest : 19 20 21 22 23, highest: 83 84 86 88 89
## ---------------------------------------------------------------------------
## VAXSAFE : VACCINES ARE SAFE  Format:AGREESCALE2F
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5    0.899    2.015   0.9882
##
## Value          1     2     3     4     5
## Frequency    144   137   112    13     4
## Proportion 0.351 0.334 0.273 0.032 0.010
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## VAXKIDS : VACCINES ARE IMPORTANT FOR CHILDREN  Format:AGREESCALE2F
```

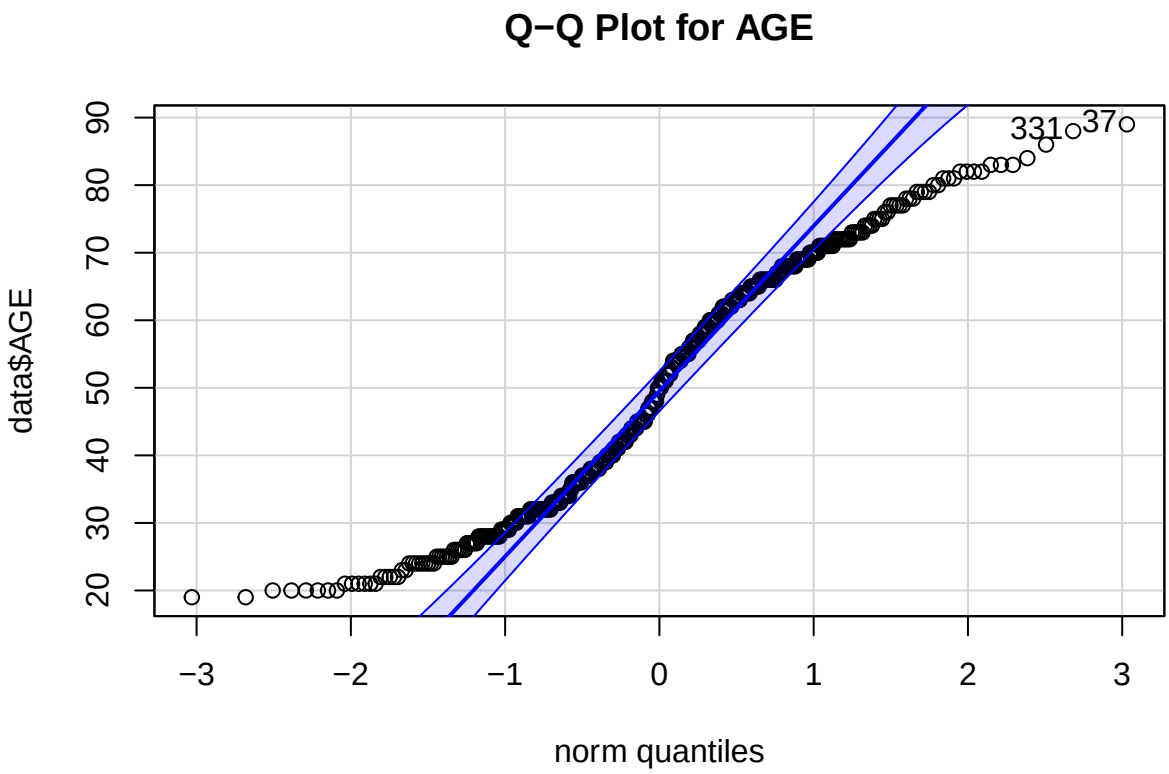
```
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5    0.823    1.673   0.8381
##
## Value          1     2     3     4     5
## Frequency    214   130    56     6     4
## Proportion 0.522 0.317 0.137 0.015 0.010
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## COVID12 : COVID VACCINE EVER  Format:YESNO
##        n  missing distinct     Info     Mean      Gmd
##      410        0        2    0.476    1.198   0.3178
##
## Value          1     2
## Frequency    329    81
## Proportion 0.802 0.198
## ----------------------------------------------------------------------------
## VAXDOHARM : VACCINES DO MORE HARM THAN GOOD  Format:AGREESCALE2C
##        n  missing distinct     Info     Mean      Gmd
##      410        0        5      0.9     3.98    1.105
##
## Value          1     2     3     4     5
## Frequency     13    18    94   124   161
## Proportion 0.032 0.044 0.229 0.302 0.393
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------

##  num [1:410] 16 14 12 13 15 14 16 14 14 18 ...
##  - attr(*, "label")= chr "Highest year of school completed"
##  - attr(*, "format.sas")= chr "EDUC"

##  Ord.factor w/ 21 levels "No Formal Schooling"<..: 17 15 13 14 16 15 17 15 15 19 ...
```




```
## # A tibble: 1 x 7
##   count  mean median  mode    sd   min   max
##   <int> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   410  14.7     14    12  2.73     2    20

##  num [1:410] 72 62 59 64 65 71 34 26 20 31 ...
##  - attr(*, "label")= chr "Age of respondent"
##  - attr(*, "format.sas")= chr "AGE"
```


Distribution of Respondents Age
Frequency
0
10
20
20
40
60
80
Age


Boxplot of Respondents' Ages
Age
80
60
40
20
−0.4
−0.2
0.0
0.2
0.4

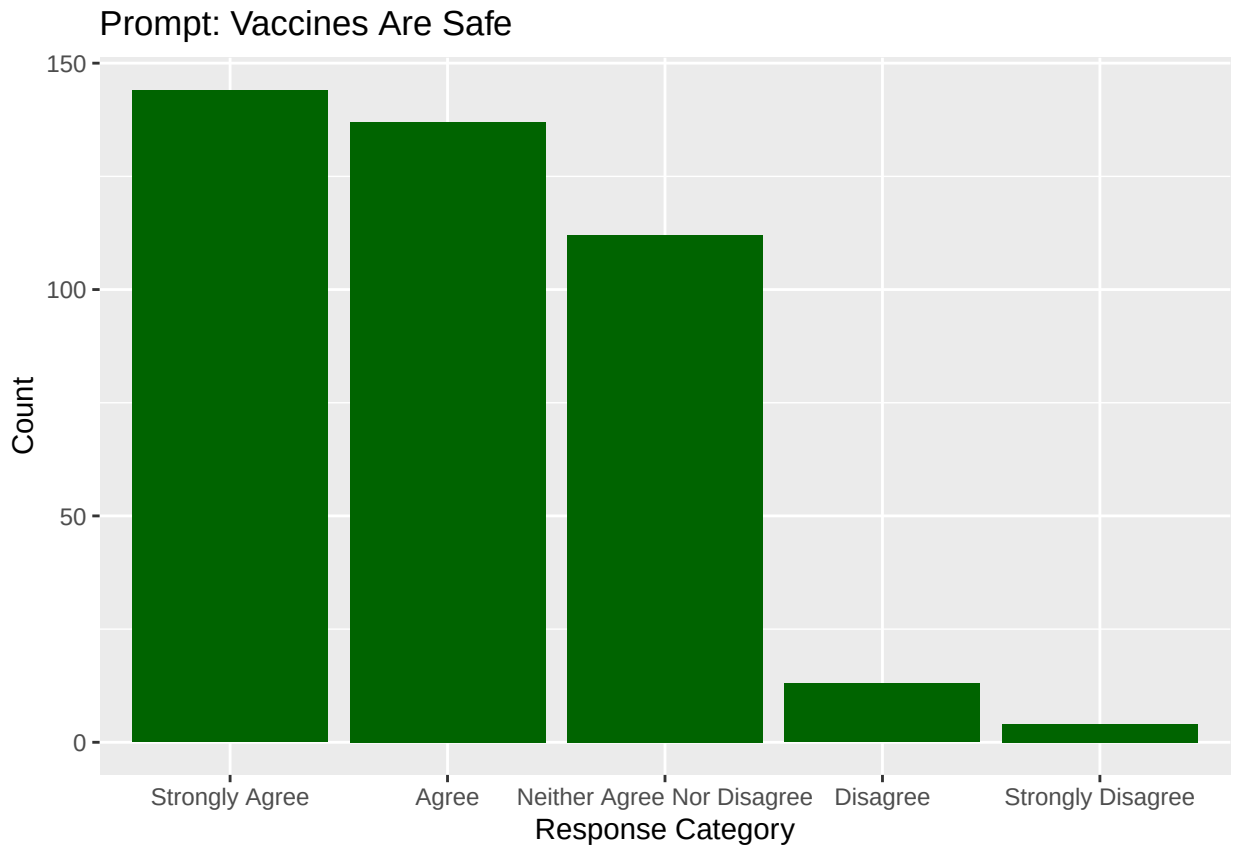



```
## [1]  37 331
```

```
## # A tibble: 1 x 7
##   count  mean median    sd   min   max  mode
##   <int> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   410  49.8     50  18.1    19    89    32
```

```
##  num [1:410] 1 4 3 1 2 1 3 1 1 1 ...
##  - attr(*, "label")= chr "VACCINES ARE SAFE"
##  - attr(*, "format.sas")= chr "AGREESCALE2F"
```

```
##  Ord.factor w/ 5 levels "Strongly Agree"<..: 1 4 3 1 2 1 3 1 1 1 ...
```

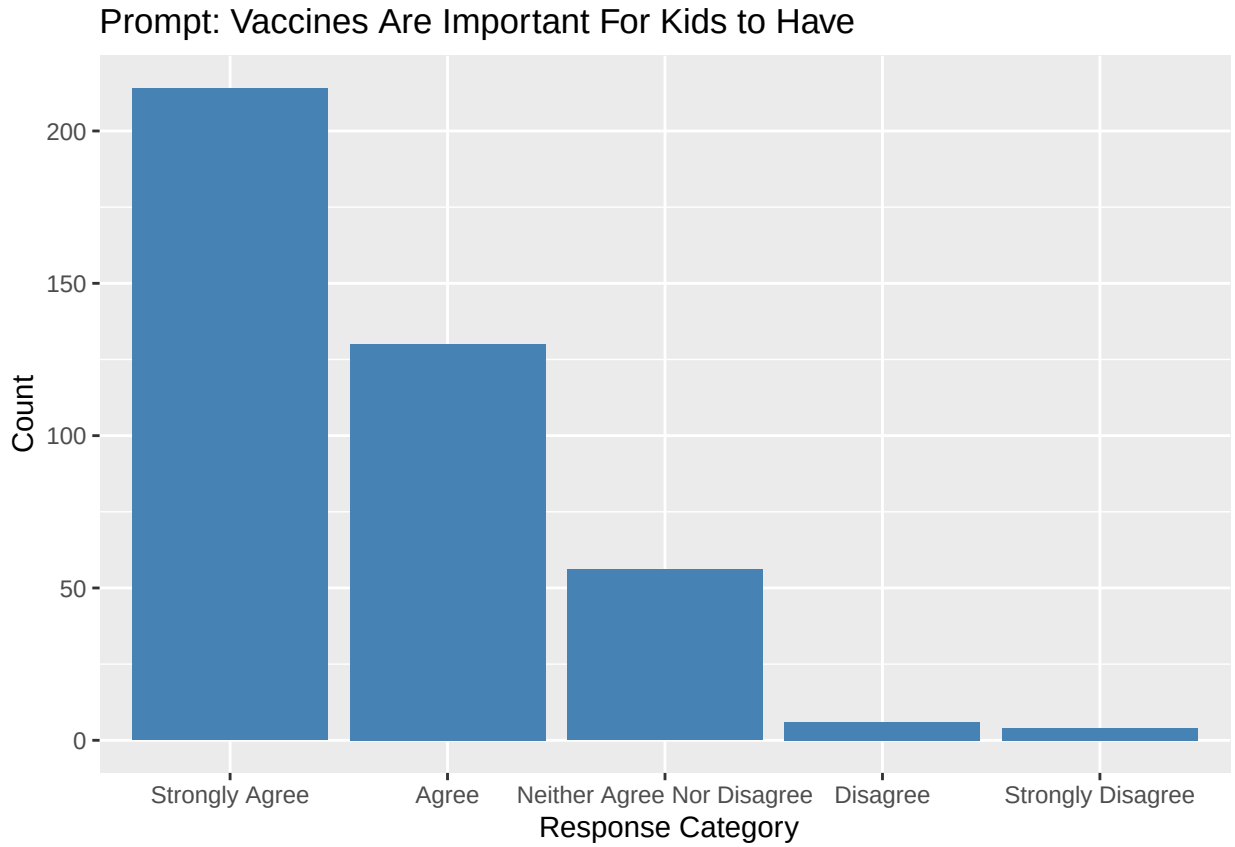



```
##
##             Strongly Agree                    Agree
##                        144                      137
## Neither Agree Nor Disagree                 Disagree
##                        112                       13
##          Strongly Disagree
##                          4
```

```
## # A tibble: 1 x 3
##   count  mean  mode
##   <int> <dbl> <dbl>
## 1   410  2.01     1
```

```
##  num [1:410] 1 3 1 1 1 2 2 1 1 1 ...
##  - attr(*, "label")= chr "VACCINES ARE IMPORTANT FOR CHILDREN"
##  - attr(*, "format.sas")= chr "AGREESCALE2F"
```

```
##  Ord.factor w/ 5 levels "Strongly Agree"<..: 1 3 1 1 1 2 2 1 1 1 ...
```

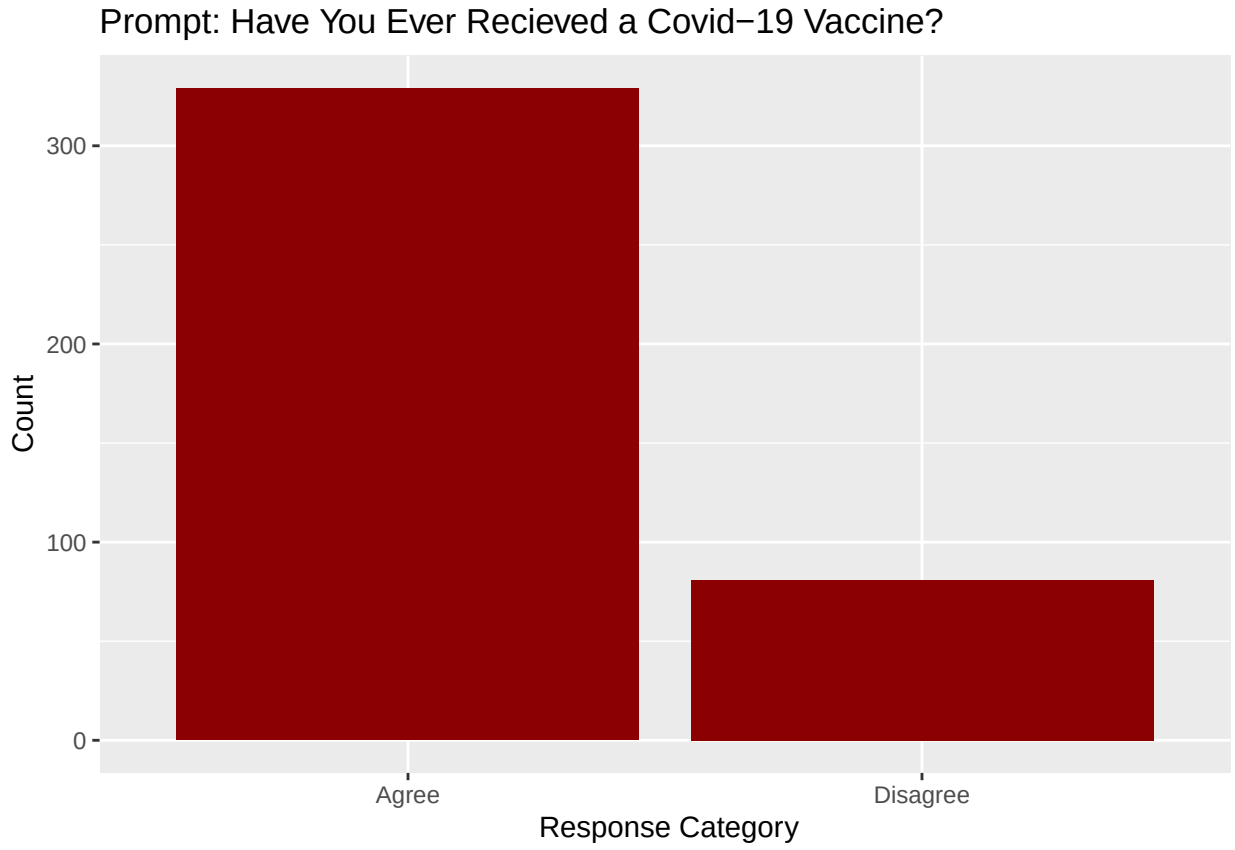



```
##
##             Strongly Agree                    Agree
##                        214                      130
## Neither Agree Nor Disagree                 Disagree
##                         56                        6
##          Strongly Disagree
##                          4
```

```
## # A tibble: 1 x 3
##   count  mean  mode
##   <int> <dbl> <dbl>
## 1   410  1.67     1
```

```
##  num [1:410] 1 2 1 1 1 1 1 1 1 1 ...
##  - attr(*, "label")= chr "COVID VACCINE EVER"
##  - attr(*, "format.sas")= chr "YESNO"
```

```
##  num [1:410] 1 0 1 1 1 1 1 1 1 1 ...
```

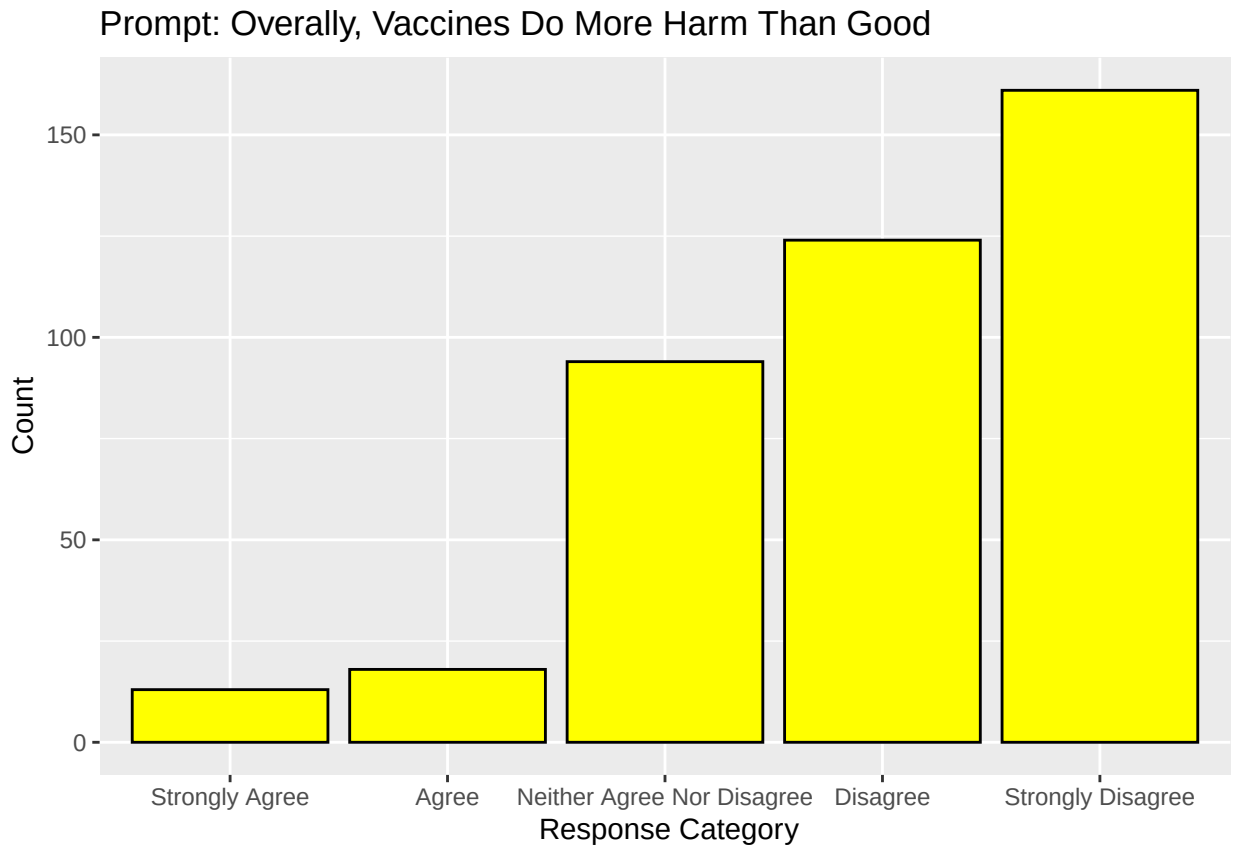



```
##
##   0   1
##  81 329
```

```
## # A tibble: 1 x 3
##   count  mean  mode
##   <int> <dbl> <dbl>
## 1   410  1.20     1
```

```
##  num [1:410] 5 2 5 4 5 5 4 3 5 5 ...
##  - attr(*, "label")= chr "VACCINES DO MORE HARM THAN GOOD"
##  - attr(*, "format.sas")= chr "AGREESCALE2C"
```

```
##  Ord.factor w/ 5 levels "Strongly Agree"<..: 5 2 5 4 5 5 4 3 5 5 ...
```




```
##
##             Strongly Agree                    Agree
##                         13                       18
## Neither Agree Nor Disagree                 Disagree
##                         94                      124
##          Strongly Disagree
##                        161
```

```
## # A tibble: 1 x 3
##   count  mean  mode
##   <int> <dbl> <dbl>
## 1   410  3.98     5
```

```
## [1] "numeric"
```

```
## [1] "ordered" "factor"
```

```
## [1] "numeric"
```

```
## [1] "ordered" "factor"
```

```
## [1] "numeric"
```

```
## [1] "ordered" "factor"
```

```
## [1] "numeric"
```

```
## Warning in cor.test.default(data$EDUC, data$VAXSAFE, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$EDUC and data$VAXSAFE
## S = 14510665, p-value = 6.307e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2632508
```

```
## [1] 0.476
```

```
## Warning in cor.test.default(data$EDUC, data$VAXDOHARM, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$EDUC and data$VAXDOHARM
## S = 7864698, p-value = 6.457e-11
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3153252
```

```
## [1] 0.498
```

```
## Warning in cor.test.default(data$AGE, data$VAXSAFE, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$AGE and data$VAXSAFE
## S = 12146886, p-value = 0.2456
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.05746793
```

```
## [1] 0.526
```

```
## Warning in cor.test.default(data$AGE, data$VAXDOHARM, method = "spearman"):
## Cannot compute exact p-value with ties
```

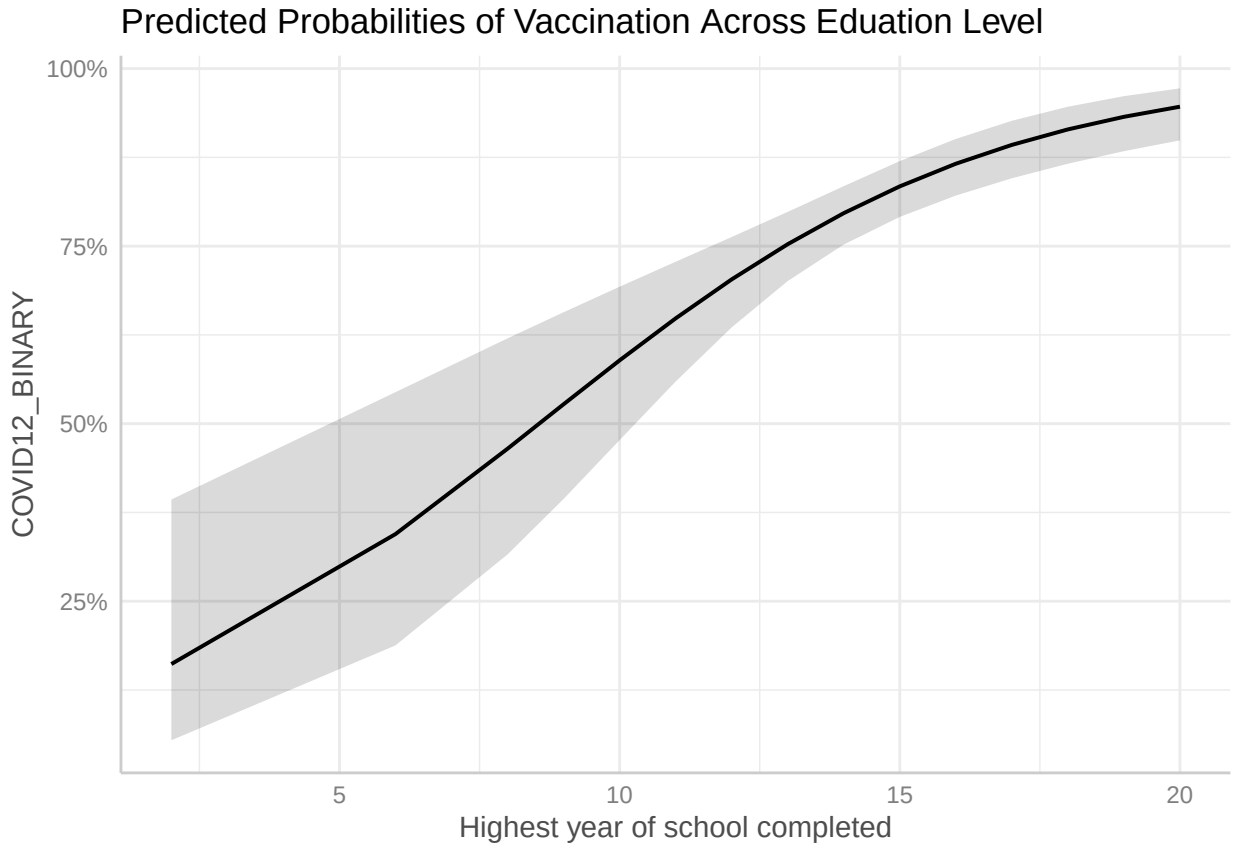
```
##
##  Spearman's rank correlation rho
##
## data:  data$AGE and data$VAXDOHARM
## S = 9860618, p-value = 0.004076
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.141567
```

```
## [1] 0.48
```

```
##               Test Spearman_Rho Spearman_P_Value Bootstrap_P_Value
## 1   EDUC & VAXSAFE  -0.26325082     6.307278e-08             0.476
## 2 EDUC & VAXDOHARM   0.31532521     6.457056e-11             0.498
## 3    AGE & VAXSAFE  -0.05746793     2.456240e-01             0.526
## 4  AGE & VAXDOHARM   0.14156701     4.075725e-03             0.480
```

```
##
## Call:
## glm(formula = COVID12_BINARY ~ EDUC, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.14881    0.72046  -2.983  0.00286 **
## EDUC         0.25094    0.05168   4.856  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.54  on 409  degrees of freedom
## Residual deviance: 380.24  on 408  degrees of freedom
## AIC: 384.24
##
## Number of Fisher Scoring iterations: 4
```

```
## (Intercept)        EDUC
##    0.116623    1.285234
```

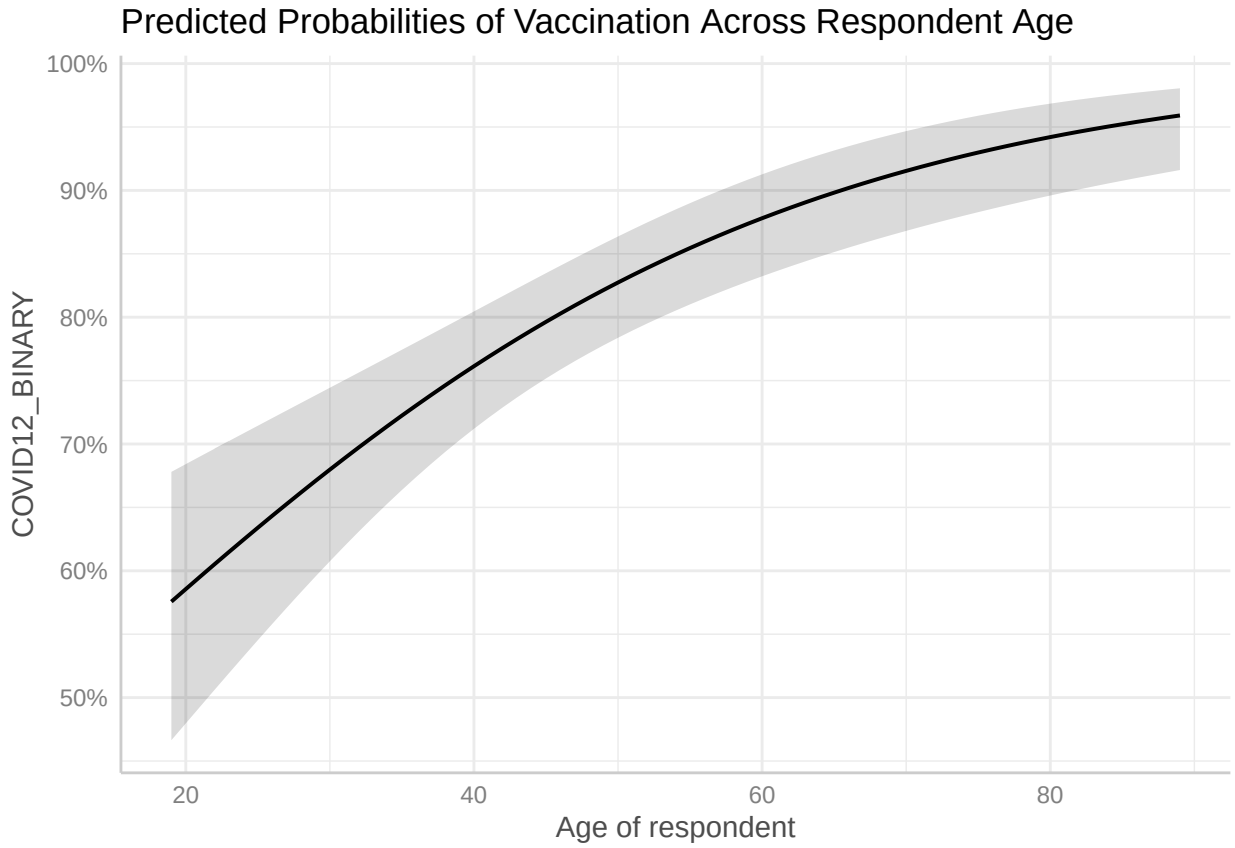



```
##  factor    AME     SE      z      p  lower  upper
##    EDUC 0.0370 0.0071 5.1937 0.0000 0.0231 0.0510
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-3.566, -0.483 )
## Calculations and Intervals on Original Scale
```

```
##
## Call:
## glm(formula = COVID12_BINARY ~ AGE, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.468435   0.357539  -1.310     0.19
## AGE          0.040710   0.007887   5.162 2.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.54  on 409  degrees of freedom
## Residual deviance: 377.01  on 408  degrees of freedom
## AIC: 381.01
##
## Number of Fisher Scoring iterations: 4

## (Intercept)         AGE
##   0.6259813   1.0415503
```




```
##  factor    AME     SE      z      p  lower  upper
##     AGE 0.0060 0.0011 5.5564 0.0000 0.0039 0.0081

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-1.1346,  0.2092 )
## Calculations and Intervals on Original Scale
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##  num [1:410] 1 4 3 1 2 1 3 1 1 1 ...
##  - attr(*, "label")= chr "VACCINES ARE SAFE"
##  - attr(*, "format.sas")= chr "AGREESCALE2F"
```

```
##  Ord.factor w/ 5 levels "Strongly Agree"<..: 1 4 3 1 2 1 3 1 1 1 ...
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = data$VAXSAFE_ORDINAL ~ EDUC + AGE, data = data)
##
## Coefficients:
##          Value Std. Error t value
## EDUC -0.181765    0.03479  -5.224
## AGE  -0.006103    0.00504  -1.211
##
## Intercepts:
##                                       Value   Std. Error t value
## Strongly Agree|Agree                 -3.6378  0.5924    -6.1412
## Agree|Neither Agree Nor Disagree     -2.1658  0.5740    -3.7730
## Neither Agree Nor Disagree|Disagree   0.2808  0.5954     0.4716
## Disagree|Strongly Disagree            1.7779  0.7352     2.4182
##
## Residual Deviance: 988.7525
## AIC: 1000.752
```

INTRODUCTION

Word Count: 585

The intersection of education and public health is an important domain, particularly in the context of vaccine attitudes. The research question this paper asks, "does education level impact people's attitudes to vaccines?" is increasingly important as the world is still adjusting to the aftermath of the COVID-19 pandemic; and any effort to understand people's reaction to medical advice and efforts is imperative to preventing something similar from happening again.

The hypothesis underpinning the research is that individuals with higher levels of education are more likely to have favourable attitudes towards vaccines. The reasoning behind this hypothesis is largely supported by existing work that associates higher education with a broader understanding of health issues, a greater capacity for critical analysis of health information, and a more pronounced trust in scientific expertise (Fowler et al., 2021; Latkin et al., 2021).

The relevance of investigating this hypothesis is multifaceted. On one level, it speaks to the recent efforts to combat the COVID-19 pandemic, where vaccine acceptance has proven pivotal to stopping the spread of the virus. On another, it addresses broader themes of trust in science and could be expanded upon in the future to research public attitudes towards health information and how socio-political factors may

play into their attitudes. The COVID-19 pandemic has also, however, highlighted the prevalence of vaccine hesitancy, even among those who are educated, challenging the notion that education is a simple fix for health misinformation (Loomba et al., 2021).

Literature Review

The existing literature demonstrates that education does generally correlate with health-promoting behaviours and positive health outcomes. For example, educated individuals are more likely to engage in preventative health behaviours and to access health services proactively (Smith et al., 2020). However, it is also important to note the nuances of this assumption by identifying how education intersects with other socio-demographic factors, like socio-economic status, to influence health behaviours (Patel et al., 2020).

Regarding vaccine attitudes, the relationship with education is changing. Prior to the pandemic, higher education was positively associated with vaccine uptake, owing to better access to health information and resources (Wilson & Wiysonge, 2020). Yet, the COVID-19 pandemic has complicated this relationship. The rapid development of vaccines, the polarised media environment, and the politicisation of health measures have led to a more fragmented relationship between education and vaccine attitudes (Sallam, 2021). Even so, Fowler et al. (2021) suggests that education facilitates a greater engagement with health systems and a nuanced understanding of risk, which could translate into positive vaccine attitudes. And to further this, Latkin et al. (2021) underscores the role of trust in science as a mediator in the relationship between education and vaccine acceptance. Education can also serve as a double-edged sword. While it can help individuals make informed health decisions, it can also equip them with the tools to rationalise vaccine scepticism, especially when mixed with ideological beliefs or mistrust in authorities (Paul et al., 2021). This is why understanding how education influences vaccine attitudes is important to understand the future of public health policy.

The significance of this question extends beyond just an academic question, this research seeks to elucidate a relationship that is vital to public health strategy in our post-pandemic lives. The World Health Organization has flagged vaccine hesitancy as a global health threat (MacDonald & SAGE Working Group on Vaccine Hesitancy, 2015). If further research is put into understanding how education along with other socio-economic and political factors impact people's perception of vaccines it would help prepare us for the next global health emergency.

METHODOLOGY

Word Count: 748

This project will analyse data through R to understand the relationship between educational attainment and vaccine attitudes. The data is sourced from the National Opinion Research Center's annual General Social Survey 2022 and the analysis functions through three main sections:

1. Exploratory and univariate analysis: Listwise deletion, variable recoding, bar charts, histogram, boxplots, QQ plot, frequency tables, summary statistics
2. Bivariate analysis: Spearman's rank correlation (Rho and p-value), bootstrapping for Spearman's (p-value), binary logistic regression (coefficients, odds ratio, AME, plot), bootstrapping for logistic regression (BCa interval)
3. Multivariate analysis: multiple ordinal logistic regression (coefficients, log odds, t-value)

Data Source and Sample

The General Social Survey (GSS) conducted by the National Opinion Research Center (NORC) in 2022 serves as the foundation for this study's data. Recognized for its extensive coverage of the United States' societal opinions, the GSS is a credible source for analysing changes in demographics and attitudes. The 2022 survey includes responses from a vast portion of the population, ensuring a representative sample of American adults.

Measures and Variables

Education (EDUC): The variable 'EDUC' quantifies respondents' educational. Recorded as integers ranging from 0 to 20, it represents a progression from 'No Formal Schooling' to '8 Years or More of College'. Ordinal.

Age (AGE): Age is measured on a continuous scale and reported in full years. The data ranges from 18 to 89, encapsulating the adult population. Given its ratio nature, age can be analysed using a variety of statistical methods, and the consistency in its reporting across NORC surveys enforces the GSS' reliability. Continuous.

Vaccine Safety (VAXSAFE): "Vaccines are safe." (GSS, 2023 p.35). Respondent's perceptions of vaccine safety are captured on a 5-point Likert scale, establishing a spectrum from 'Strongly Agree' to 'Strongly Disagree'. Ordinal.

Vaccination for Children (VAXKIDS): "Vaccines are important for children to have." (GSS, 2023 p.35). Like VAXSAFE, VAXKIDS uses a 5-point Likert scale to measure agreement with the importance of vaccinating children. The ordinal nature of this variable allows for relative comparisons between levels of agreement or disagreement. Ordinal.

COVID-19 Vaccination (COVID12): "Have you ever received a COVID-19 vaccine?" (GSS, 2023 p35). This binary variable records whether respondents have received a COVID-19 vaccine. Binary.

Vaccines and Harm (VAXDOHARM): "Overall, vaccinations do more harm than good." (GSS, 2023 p.29). Respondent's opinion on whether vaccines do more harm than good are also measured on a 5-point Likert scale, providing an ordinal variable that gauges the level of agreement with the statement. Ordinal.

Data Cleaning and Preprocessing

Initially, a considerable number of missing values were identified across variables. Traditional imputation techniques such as mean substitution, KNN testing, and multiple imputation were considered. But, due to the volume of missing data — nearing 98% for some variables — these techniques were deemed inappropriate. Imputing these data could introduce significant biases and distort the genuine patterns in the dataset. Additionally, as the data was originally downloaded as SAS documentation, variable recoding was necessary to change the variables from SAS documentation to ones appropriate for analysis in R, like ordered factors.

Statistical Analysis

For *univariate analysis*, Histograms and boxplots visualized the distribution and identified any potential outliers or skewness in AGE. For ordinal variables, bar charts depicted the frequency of each category, and measures of central tendency were calculated to provide a summary of the responses.

For *bivariate analysis,* Spearman's rank correlation was used to assess the association between education and vaccine-related attitudes due to their ordinal nature. Given the binary outcome of the COVID-19 vaccination variable, logistic regression was the most suitable method for modelling dichotomous data (James et al., 2022).

Lastly, for *multivariate analysis* multiple ordinal logistic regression (OLR) was the chosen method as it is important to use models that account for the ordered nature of the response variable (Hastie et al., Chapter 4). Bootstrap methods were incorporated to estimate the sampling distribution of the Spearman correlation and logistic regression coefficients, providing a non-parametric approach to create confidence intervals (James et al., 2021). This was particularly important given the reduced sample size post-listwise deletion. The bootstrap approach, using 500 and sometimes 1000 replications, allowed for the assessment of the stability and robustness of the findings

Justification of Methodological Choices

The methodological choices are underpinned by the nature of the data and the research questions posed. Logistic regression's ability to deal with dichotomous outcomes, Spearman's correlation's suitability for ordinal data, and OLR's capacity to handle ordered categorical responses, all align with the data types present in my subset of the GSS dataset.

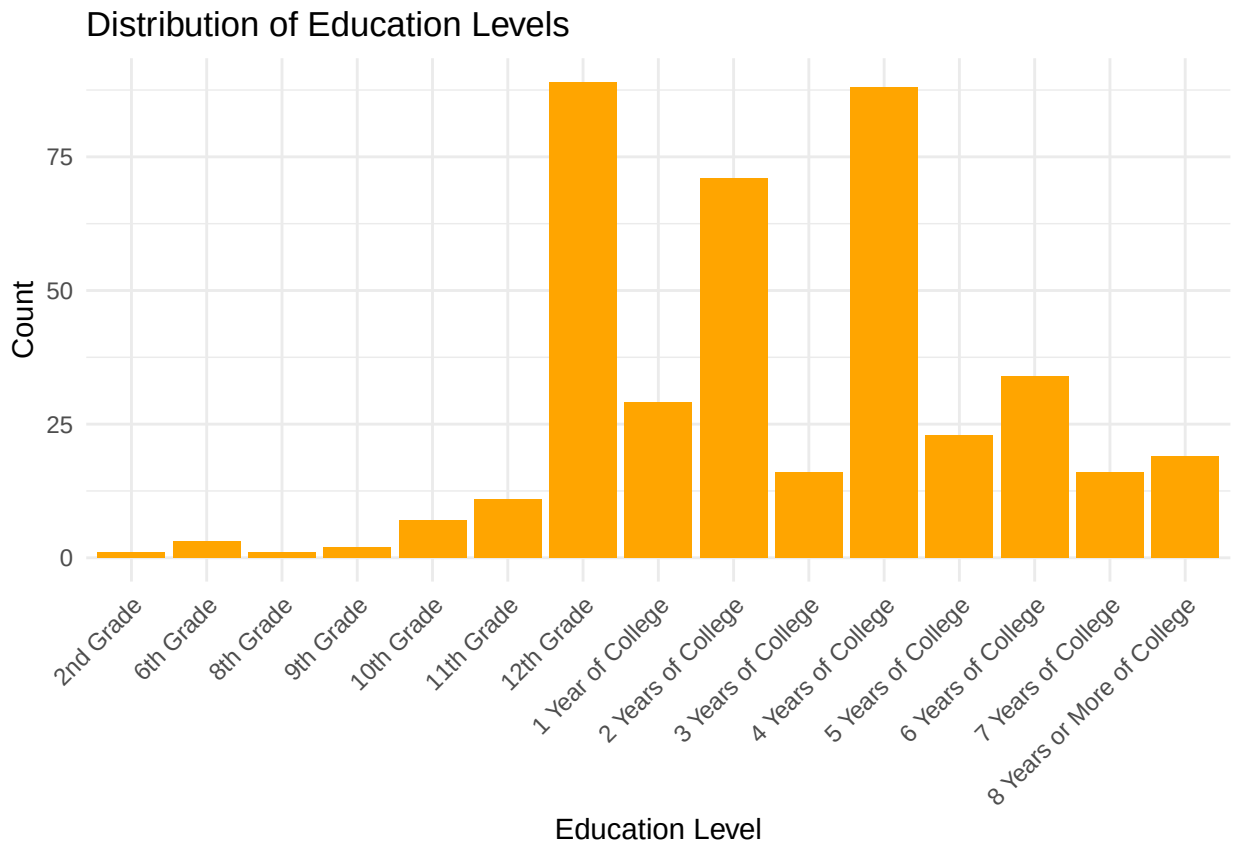
FINDINGS AND CONCLUSION

Word Count: 990

Univariate Analysis

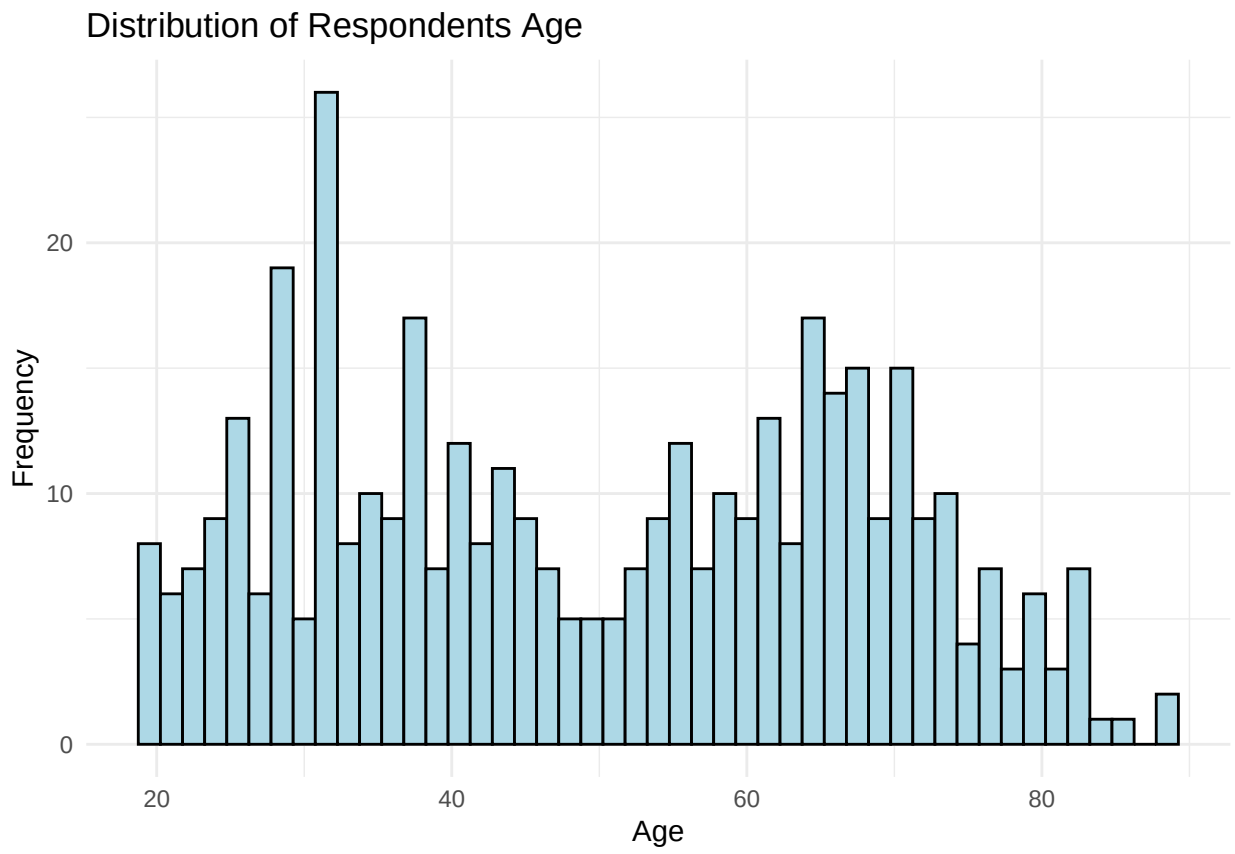
EDUC (Education): The dataset revealed a wide range of educational attainment, from no formal schooling to advanced college degrees. The most common response was high school completion (12th grade). A bar chart (Graph: Distribution of Education Levels) visually depicted this distribution, highlighting the skew towards higher education levels.

```
##  Ord.factor w/ 21 levels "No Formal Schooling"<..: 17 15 13 14 16 15 17 15 15 19 ...
```



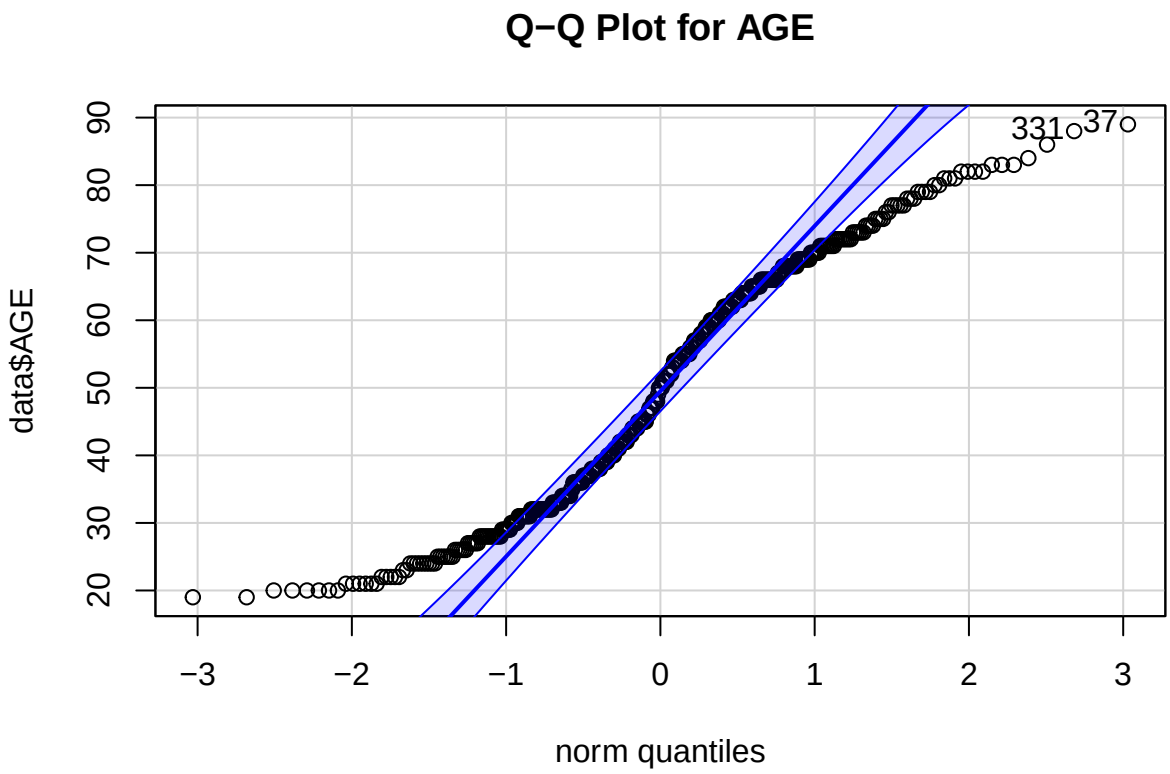
AGE (Age): The age of respondents varied from 18 to 89, with a mean age of approximately 50 years (49.8). The distribution showed a relatively even spread across different age groups, as indicated by a histogram (Graph: Distribution of Respondents' Age).

```
#Histogram for AGE
ggplot(data, aes(x = AGE)) +
  geom_histogram(binwidth = 1.5, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Respondents Age", x = "Age", y = "Frequency") +
  theme_minimal()
```




```r
#boxplot for AGE
ggplot(data, aes(y = AGE)) +
  geom_boxplot() +
  labs(title = "Boxplot of Respondents' Ages", y = "Age")
```

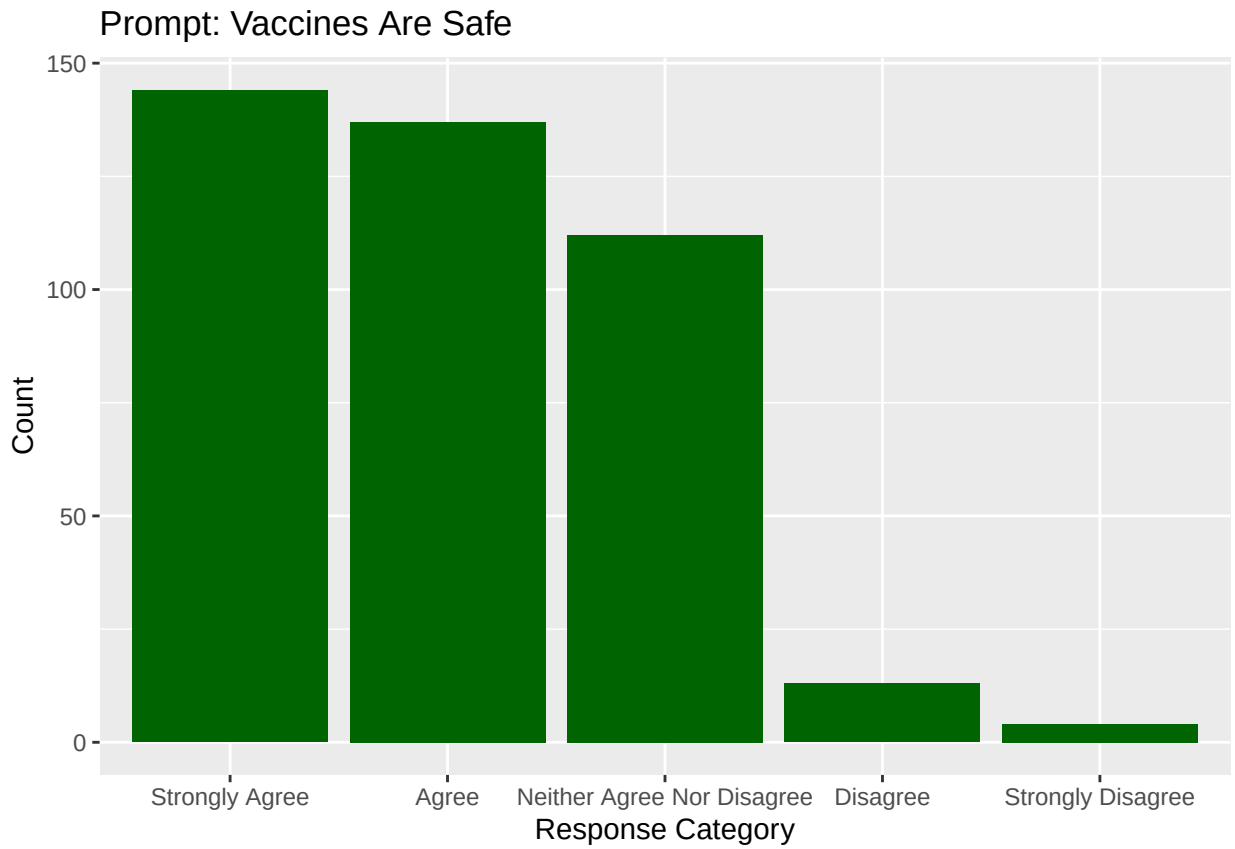

```
#Q-Q plot for AGE
qqPlot(data$AGE, main = "Q-Q Plot for AGE")
```




```
## [1]  37 331
```

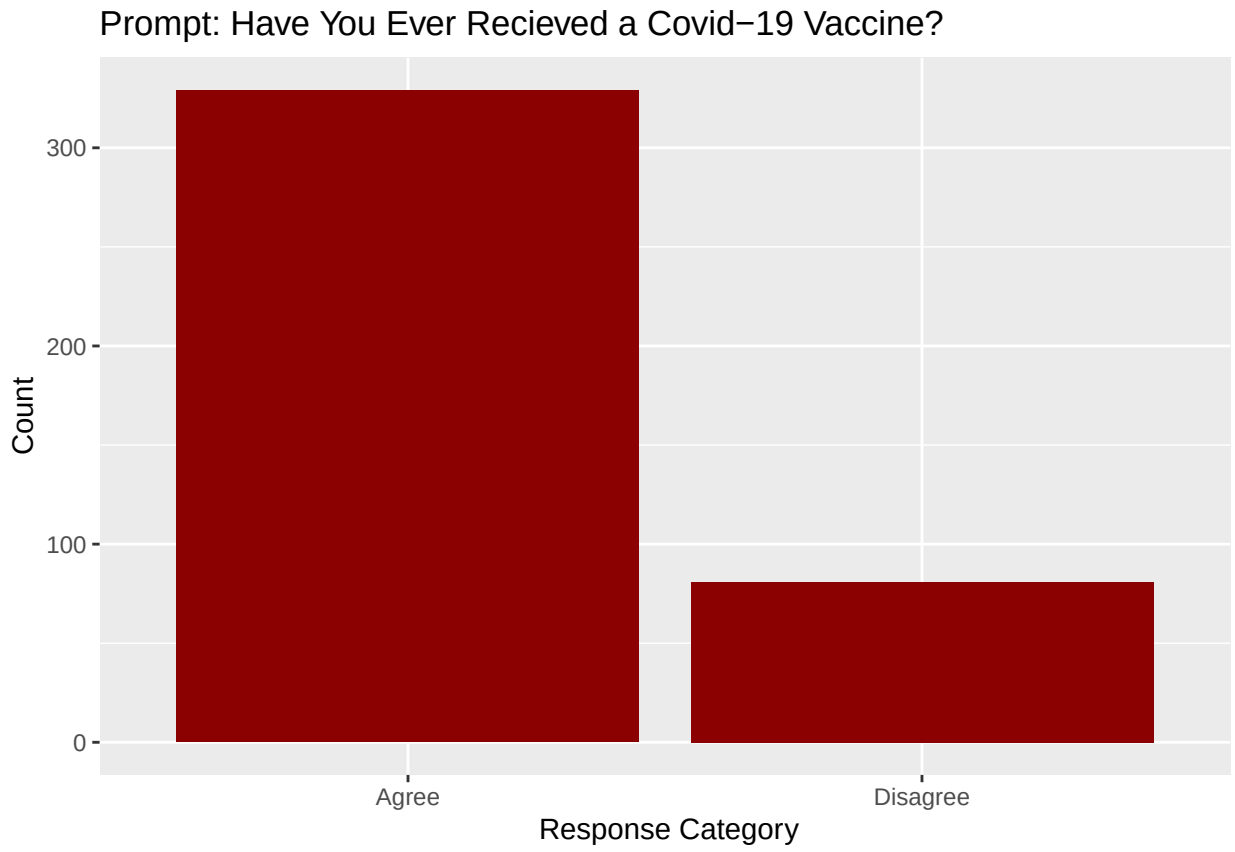
VAXSAFE (Attitudes Towards Vaccine Safety): Most respondents tended to agree that vaccines are safe, with a smaller proportion expressing strong disagreement. The distribution of the bar chart suggests a generally positive attitude toward vaccine safety among the participants.

```
#barchart for VAXSAFE_LABEL
ggplot(data, aes(x = factor(VAXSAFE_ORDINAL))) +
  geom_bar(fill = "darkgreen") +
  labs(title = "Prompt: Vaccines Are Safe", x = "Response Category", y = "Count")
```



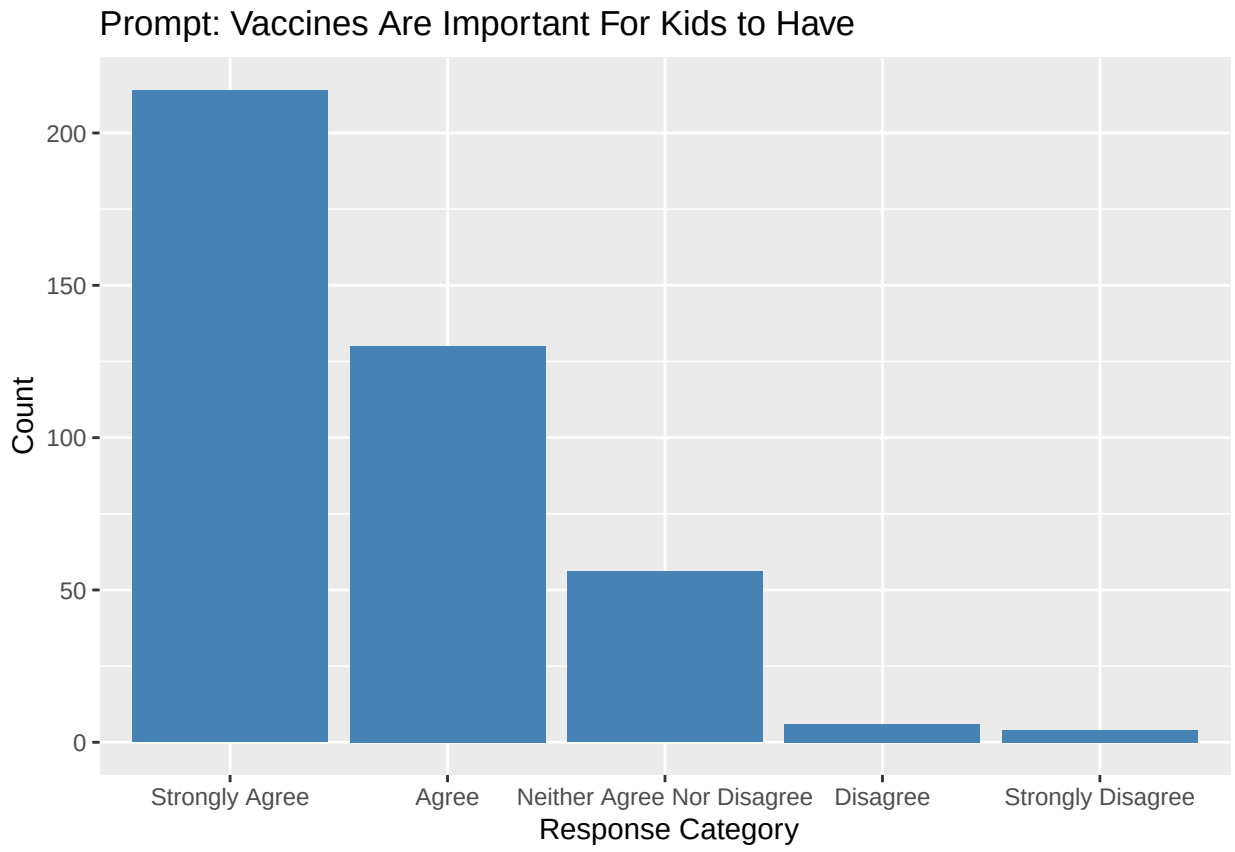
COVID12 (COVID-19 vaccination status): A significant majority (about 80%) reported receiving a COVID-19 vaccine.

```
#barchart for COVID12
ggplot(data, aes(x = factor(COVID12_LABELS))) +
  geom_bar(fill = "darkred") +
  labs(title = "Prompt: Have You Ever Recieved a Covid-19 Vaccine?", x = " Response Category", y = "Cou
```

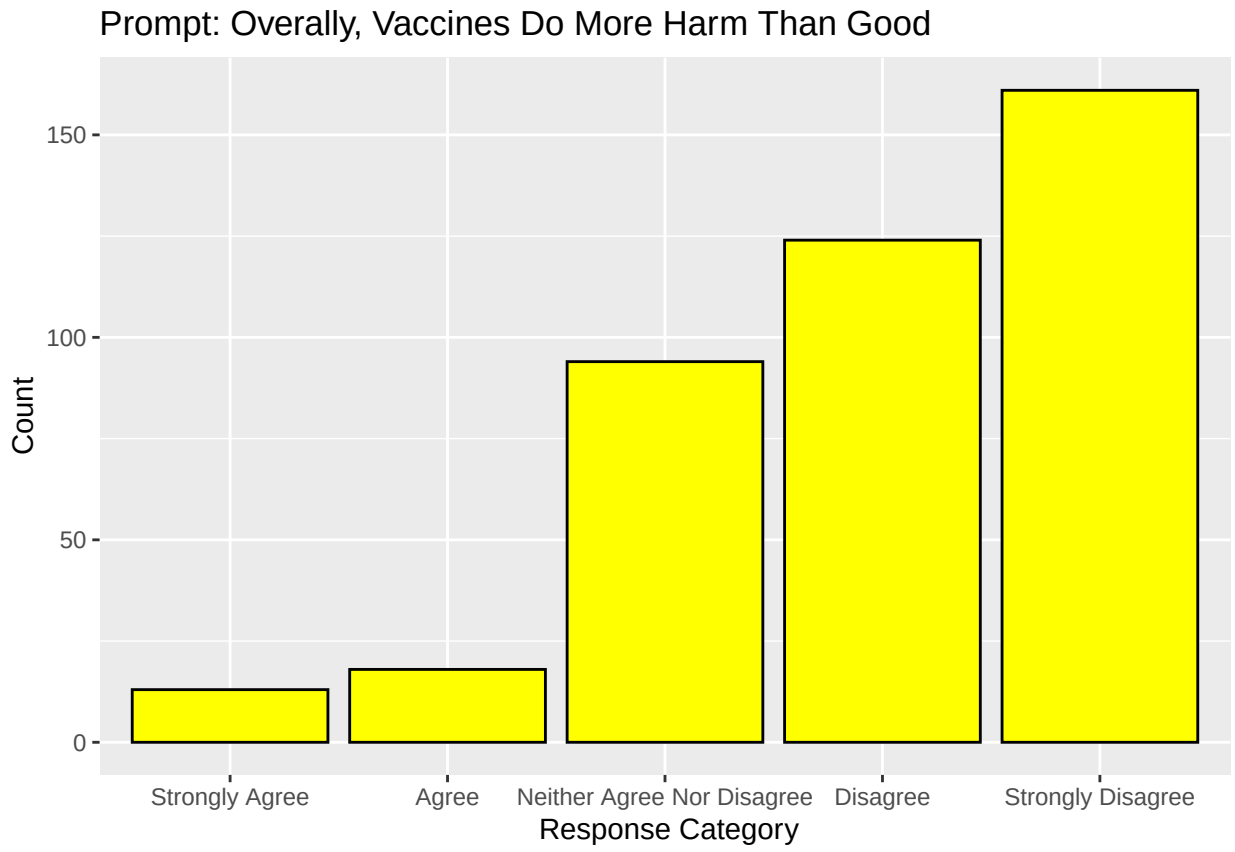


VAXKIDS, VAXDOHARM: There was a general trend towards agreement that vaccines are safe and important for children, and towards vaccines not being harmful. However, VAXDOHARM and VAXKIDS are analysed less as VAXSAFE is prioritized for most of the following analyses.

```
#barchart for VAXKIDS_ORDINAL
ggplot(data, aes(x = factor(VAXKIDS_ORDINAL))) +
  geom_bar(fill = "steelblue") +
  labs(title = "Prompt: Vaccines Are Important For Kids to Have", x = " Response Category", y = "Count")
```




```
#barchart for VAXDOHARM
ggplot(data, aes(x = factor(VAXDOHARM_ORDINAL))) +
  geom_bar(fill = "yellow", color = "black") +
  labs(title = "Prompt: Overally, Vaccines Do More Harm Than Good", x = " Response Category", y = "Coun
```



Bivariate Analysis

1. Education and Vaccine Safety (VAXSAFE):

   Method: Spearman's rank correlation coefficient, suitable for ordinal data. The below equation where d = difference in ranks of corresponding values, n = number of observations.

Equation:

$$\rho = 1 - (6\Sigma d_i^2)/(n^3 - n)$$

Spearman's correlation indicated a weak negative correlation between education level and belief in vaccine safety (rho = -0.263, $p < 0.001$). VAXSAFE is ordinal ranging 1-5 with 1 being strongly agree, and 5 being strongly disagree. Therefor the negative relationship implies as EDUC increases VAXSAFE rank decrease, or, as education level increases people are more likely to agree that vaccines are safe.

The weak correlation coefficient indicates that while the relationship is statistically significant, the strength of association is minimal.

2. Education and Belief in Vaccine Harm (VAXDOHARM):

   Method: Spearman's rank correlation coefficient.

   Analysis: The Correlation Coefficient: 0.315, this represents a moderate positive association, that as EDUC increases VAXDOHARM responses also increase (1-strongly agree, 5-strongly disagree). So as education increases, respondents are more likely to disagree that vaccines do harm. The P-value was 6.457e-11, this strongly suggests that the correlation is statistically significant as it is well below the accepted value of $P < .05$.

```
results_df <- data.frame(
  Test = c("EDUC & VAXSAFE", "EDUC & VAXDOHARM", "AGE & VAXSAFE", "AGE & VAXDOHARM"),
  Spearman_Rho = c(correlation_educ_vaxsafe$estimate, correlation_educ_vaxdoharm$estimate,
                   correlation_age_vaxsafe$estimate, correlation_age_vaxdoharm$estimate),
  Spearman_P_Value = c(correlation_educ_vaxsafe$p.value, correlation_educ_vaxdoharm$p.value,
                       correlation_age_vaxsafe$p.value, correlation_age_vaxdoharm$p.value),
  Bootstrap_P_Value = c(boot_p_value_educ_vaxsafe, boot_p_value_educ_vaxdoharm,
                        boot_p_value_age_vaxsafe, boot_p_value_age_vaxdoharm))
```

```
print(results_df)
```

```
##               Test Spearman_Rho Spearman_P_Value Bootstrap_P_Value
## 1   EDUC & VAXSAFE  -0.26325082     6.307278e-08             0.476
## 2 EDUC & VAXDOHARM   0.31532521     6.457056e-11             0.498
## 3    AGE & VAXSAFE  -0.05746793     2.456240e-01             0.526
## 4  AGE & VAXDOHARM   0.14156701     4.075725e-03             0.480
```

3. Education and Vaccination status (COVID12):

   Model: Binary logistic regression with COVID12_BINARY (whether or not an individual is vaccinated against COVID-19) as the response variable and EDUC (education level) as the predictor.

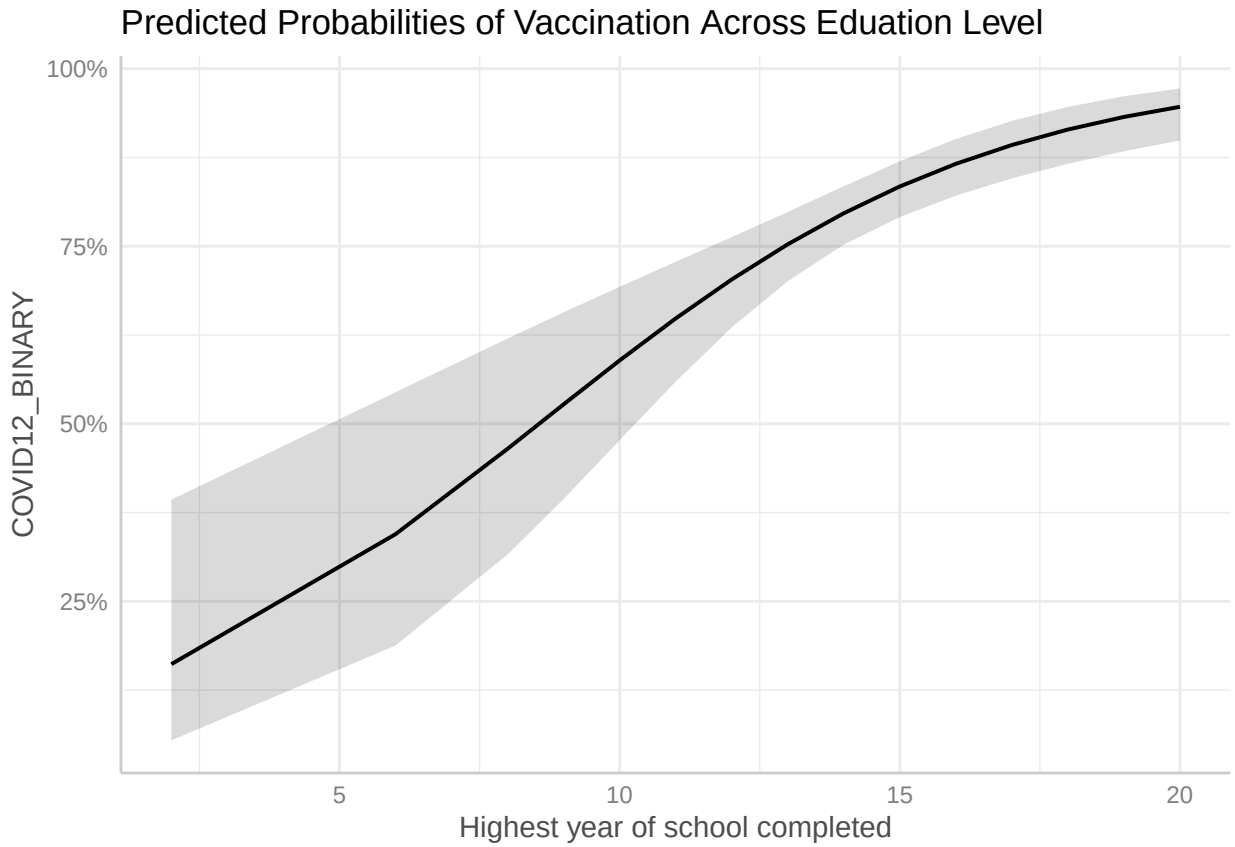
```
    covid_model <- glm(COVID12_BINARY ~ EDUC, family = binomial(link = "logit"), data = data)
    summary(covid_model)# coeffcient ofIntercept 2.14881 = negative log-odds of
```

```
##
## Call:
## glm(formula = COVID12_BINARY ~ EDUC, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.14881    0.72046  -2.983  0.00286 **
## EDUC         0.25094    0.05168   4.856  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.54  on 409  degrees of freedom
## Residual deviance: 380.24  on 408  degrees of freedom
## AIC: 384.24
##
## Number of Fisher Scoring iterations: 4
```

```
    #calculate and interpret odds ratios
    exp(covid_model$coefficients)
```

```
## (Intercept)        EDUC
##    0.116623    1.285234
```

```r
#predicted probabilities and marginal effects
ggpredict(covid_model, terms = "EDUC[all]") %>%
  plot() +
  labs(title = "Predicted Probabilities of Vaccination Across Eduation Level")
```



Equation:

$$log(p/(1-p)\ ) = *\beta * 0\ + *\beta * 1\ * X * 1\ + *\beta * 2\ * X * 2\ + ... + *\beta n *\ \ * Xn*$$

Outcome:

$$log(p/(1 - *p) *\ \ ) = -2.14881$$

Coefficients:

Intercept – The coefficient of the intercept is -2.14881. This shows that the odds are less than 1 (because the log of a number between 0 and 1 is negative), which suggests lower odds of being vaccinated at the baseline education level (minimal education).

EDUC – The coefficient for EDUC is 0.25094. This positive coefficient suggests that as the education level increases, the log odds of being vaccinated also increase.

Statistical Significance: Both the intercept and the EDUC coefficient have p-values well below 0.05, indicating that they are statistically significant.

Odds Ratio: The odds ratio for EDUC is approximately 1.285, meaning that with each additional unit increase in education, the odds of being vaccinated increase by roughly 29% (28.5%).

graph

Average Marginal Effects (AME): The AME of 0.0370 implies that on average, each additional unit of education increases the probability of being vaccinated by about 3.7%.

```
covid_model <- glm(COVID12_BINARY ~ EDUC, family = binomial(link = "logit"), data = data)
summary(covid_model)# coeffcient ofIntercept 2.14881 = negative log-odds of being agree when EDUC at ba
```

```
##
## Call:
## glm(formula = COVID12_BINARY ~ EDUC, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.14881    0.72046  -2.983  0.00286 **
## EDUC         0.25094    0.05168   4.856  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.54  on 409  degrees of freedom
## Residual deviance: 380.24  on 408  degrees of freedom
## AIC: 384.24
##
## Number of Fisher Scoring iterations: 4
```

Bootstrap confidence Intervals: The BCa (bias-corrected and accelerated) bootstrap confidence interval is (-3.566, -0.483). The wide range suggests we should be cautious interpreting the exact effect size, especially considering the potential for overfitting considering how small the dataset is.

```
print(boot_ci)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-1.1346,  0.2092 )
## Calculations and Intervals on Original Scale
```

Multivariate Analysis

1. Ordinal Logistic Regression (VAXSAFE_ORDINAL):

   The model indicated a negative relationship between education/age and higher categories of agreement (1-strongly agree, 5- strongly disagree), as education and age increase respondents are more likely to agree that vaccines are safe.

Model: OLR, or proportional odds model, is used here because the response variable ("VAXSAFE_ORDINAL") is ordinal. This model assumes that the relationship between each pair of outcome categories is the same.

Coefficients:

EDUC – The coefficient for EDUC is -0.182, with a highly significant t-value of -5.224. This indicates that as the level of education increases, the log-odds of having a higher level of agreement that vaccines are safe decreases.

AGE – The coefficient for AGE is -0.006103, the negative coefficient implies that older respondents are more likely to agree that vaccines are safe. This relationship, however, is not as statistically significant as VAXSAFE's relationship with education.

Model Fit: The high residual deviance (988 degrees of freedom) and AIC suggest that the model may not be the best fit for the data, likely due to the small sample size.

```
#multivariate ordinal logistic regression
olr_model <- polr(data$VAXSAFE_ORDINAL ~ EDUC + AGE, data = data)
#check the summary
summary(olr_model)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = data$VAXSAFE_ORDINAL ~ EDUC + AGE, data = data)
##
## Coefficients:
##          Value Std. Error t value
## EDUC -0.181765    0.03479  -5.224
## AGE  -0.006103    0.00504  -1.211
##
## Intercepts:
##                                     Value   Std. Error t value
## Strongly Agree|Agree                -3.6378  0.5924    -6.1412
## Agree|Neither Agree Nor Disagree    -2.1658  0.5740    -3.7730
## Neither Agree Nor Disagree|Disagree  0.2808  0.5954     0.4716
## Disagree|Strongly Disagree           1.7779  0.7352     2.4182
##
## Residual Deviance: 988.7525
## AIC: 1000.752
```

Conclusion

The analysis reveals that higher education is associated with more favourable opinions about vaccine safety, aligning with higher vaccination rates among the educated. This indicates a direct relationship between education and positive health beliefs and supports my hypothesis.

Cautions: The high rate of missing data and the resulting reduction in sample size necessitate caution. The findings might not be fully representative of the broader population. Additionally, the bootstrapped p-values indicate that some of the observed correlations are not as robust as initially thought.

Future Research: Given more time and resources, a more comprehensive dataset with fewer missing values would be ideal. Additionally, employing different imputation methods for missing data could provide a more nuanced understanding of the relationship. Also, I would have gone back and recoded the ordinal variables

such that they are more intuitive for interpretation as I thought some of the models were returning results that supported H0 due to the unintuitive coding of VAXSAFE.

My hypothesis that higher education leads to more favourable views on vaccines is supported by bivariate and multivariate analysis, but more research should be conducted with a larger dataset for truly robust research on the topic.

BIBLIOGRAPHY

Fowler, F. J., Brown, J. A., & Weiland, A. J. (2021). The effect of education on health behaviour changes after a diabetes diagnosis among older Americans. Journal of Aging and Health, 33(7-8), 511-521.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R. Second Edition.

Latkin, C. A., Dayton, L., Yi, G., Konstantopoulos, A., & Boodram, B. (2021). Trust in a COVID-19 vaccine in the U.S.: A social-ecological perspective. Social Science & Medicine, 270, 113684.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nature Human Behaviour, 5(3), 337-348.

MacDonald, N. E., & SAGE Working Group on Vaccine Hesitancy. (2015). Vaccine hesitancy: Definition, scope and determinants. Vaccine, 33(34), 4161-4164.

Sallam, M. (2021). COVID-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. Vaccines, 9(2), 160.