DisclosureRisk

Kevin Ros

2/16/2020

```
library(ggplot2)
library(coda)
library(runjags)
library(fastDummies)
data = data.frame(read.csv("personsxcsv/personsx.csv",header=TRUE))
```

The NIHS Person data set is :

NHIS collects data on both adult and children's mental health and mental disorders. For adults, this includes serious psychological distress and feelings of depression and anxiety. For children, this includes the presence of attention deficit/hyperactivity disorder and autism spectrum disorder. The NHIS also examines mental health service use and whether individuals have unmet mental health needs. Questions about recent anxiety or frequent stress have been included in previous years.

Note that much of this data is sensitive, as it pertains to health, which may be protected by HIPAA.

Two things I learned from this dataset/exercise:

(1) Many entries have NA for multiple variables, so the knowledge (or lack thereof) of a characteristic of someone may significantly increase disclosure / identification risk.

(2) The existance of "Universes" where certain variables are bounded to the value of another, broader variable. See PDF!

The number of male (1) and female(2) records are approx. equal, so not much identification disclosure risk.

```
nrow(data[data$SEX == 1,]) # Male
```

```
## [1] 35549
```

```
nrow(data[data$SEX == 2,]) # Female
```

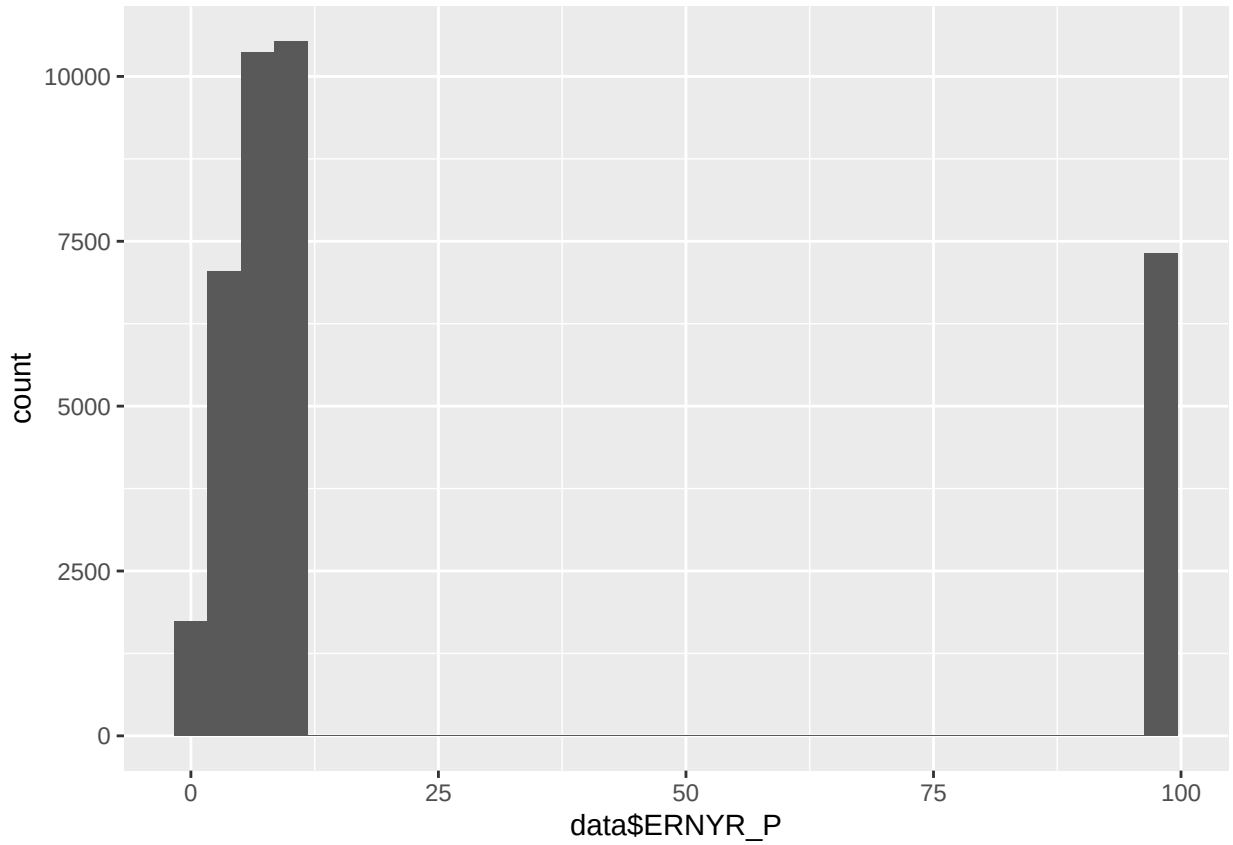
```
## [1] 37282
```

Looks like ~35000 out of ~72000 do not have records for yearly earnings.

```
ggplot(data, aes(x = data$ERNYR_P)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 35847 rows containing non-finite values (stat_bin).
```

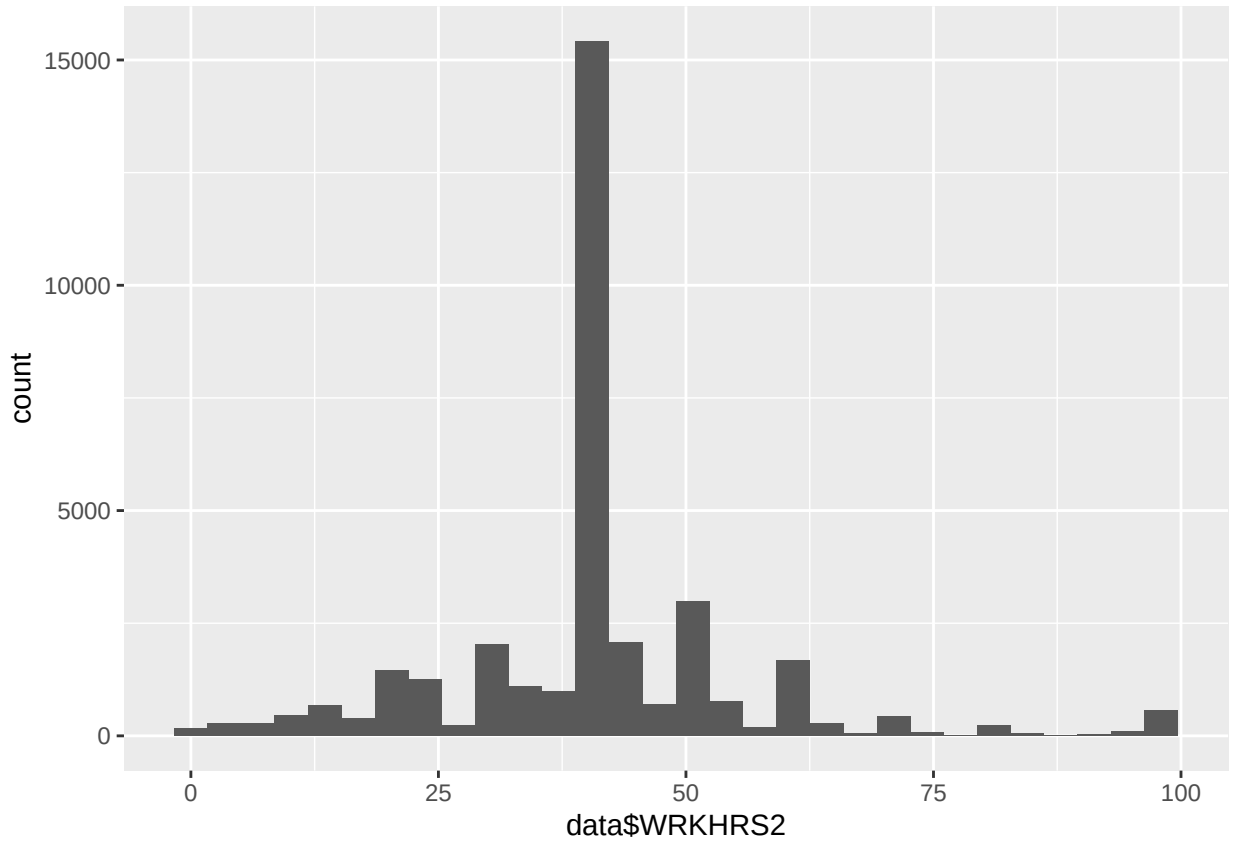


WRKHRS2 records the number of hours a participant has worked in the past week. 40 is the overwhelming response, but anything other than 40 is a potential disclosure risk. Additionally, ~37000 participants did not answer this question.

```
ggplot(data, aes(x = data$WRKHRS2)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

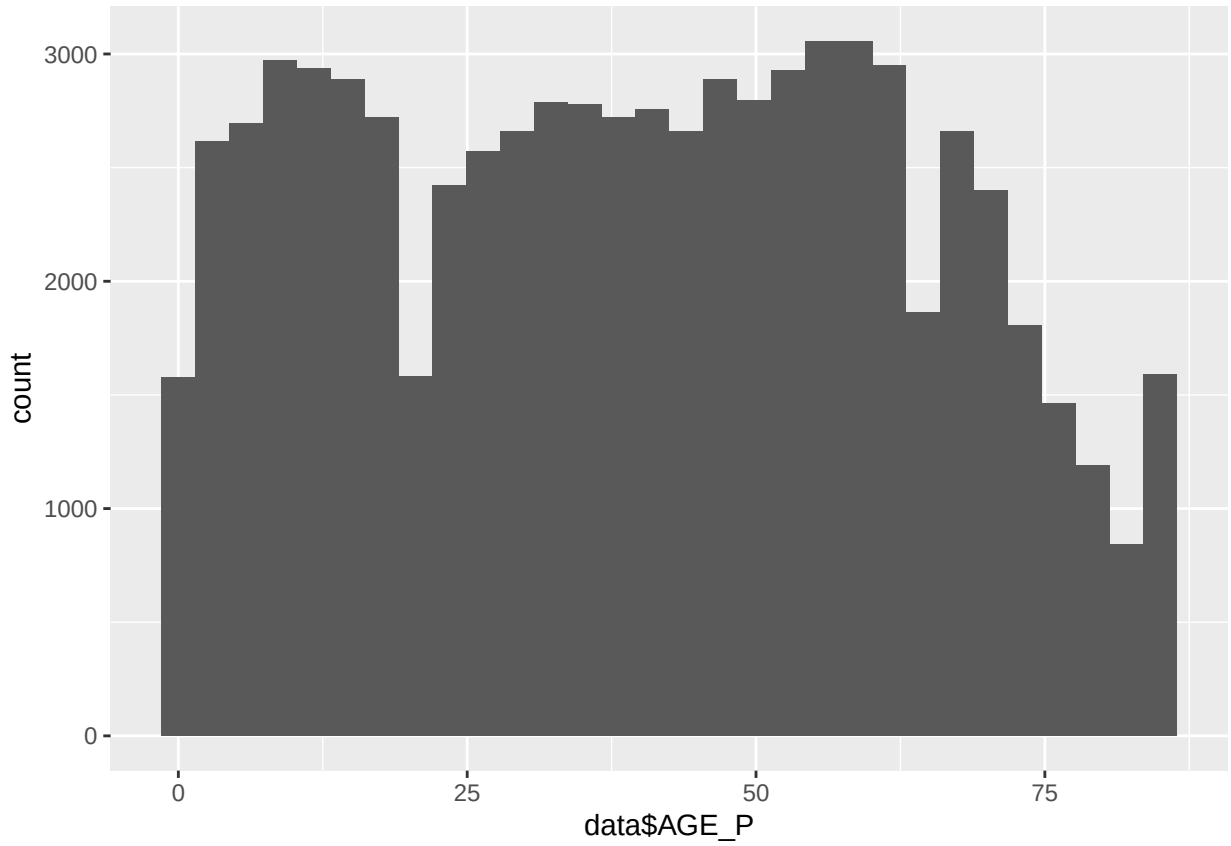
```
## Warning: Removed 37910 rows containing non-finite values (stat_bin).
```



The age of everyone in the survey. It seems like all participants answered this quesiton.

```
ggplot(data, aes(x = data$AGE_P)) + geom_histogram()
```

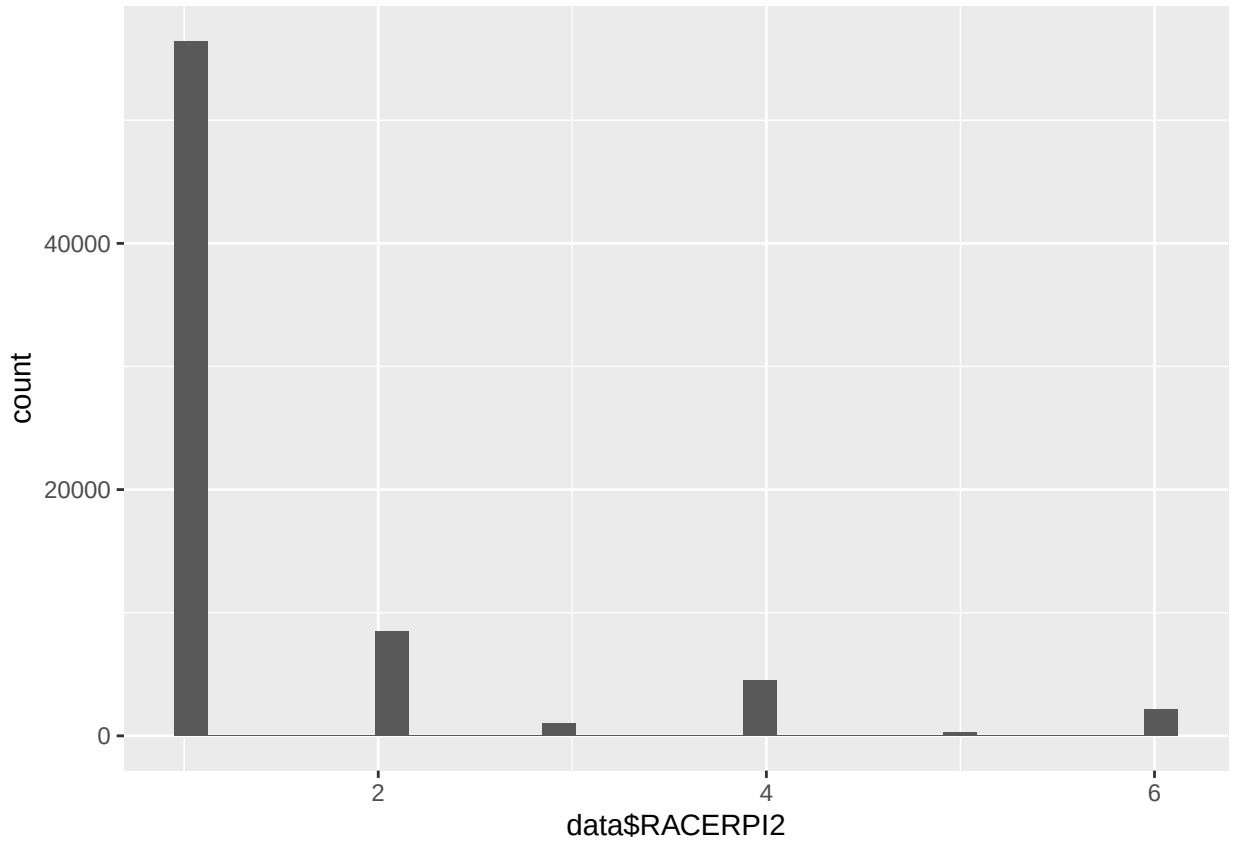
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The races of everyone in the survey.

```
ggplot(data, aes(x = data$RACERPI2)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



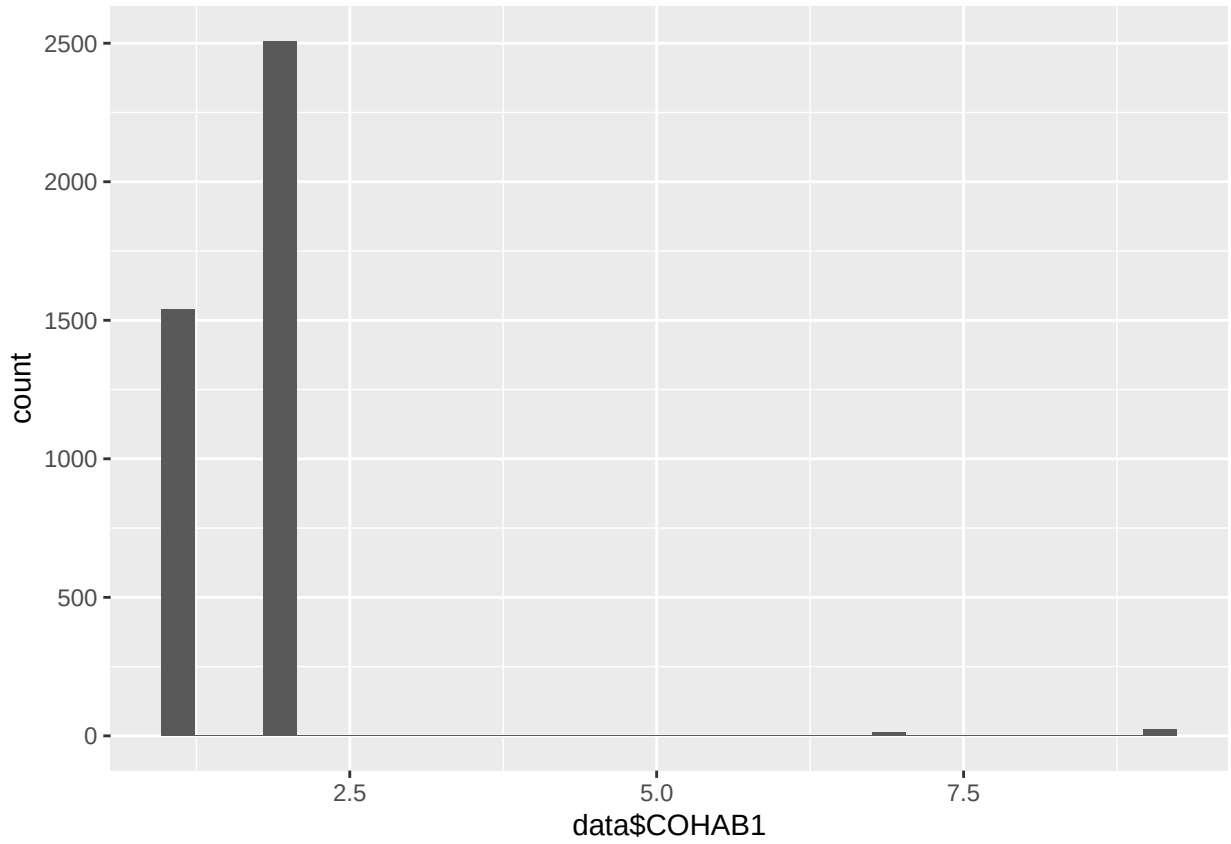

```
#01 White only
#02 Black/African American only
#03 AIAN only
#04 Asian only
#05 Race group not releasable*
#06 Multiple race
```

"Has the person ever been married?" where 1 = yes, 2 = no Universe = COHAB1 = 1, "What is the person's current legal marital status?" 1 Married 2 Widowed 3 Divorced 4 Separated 7 Refused 8 Not ascertained 9 Don't know

```
ggplot(data, aes(x = data$COHAB1)) + geom_histogram()
```

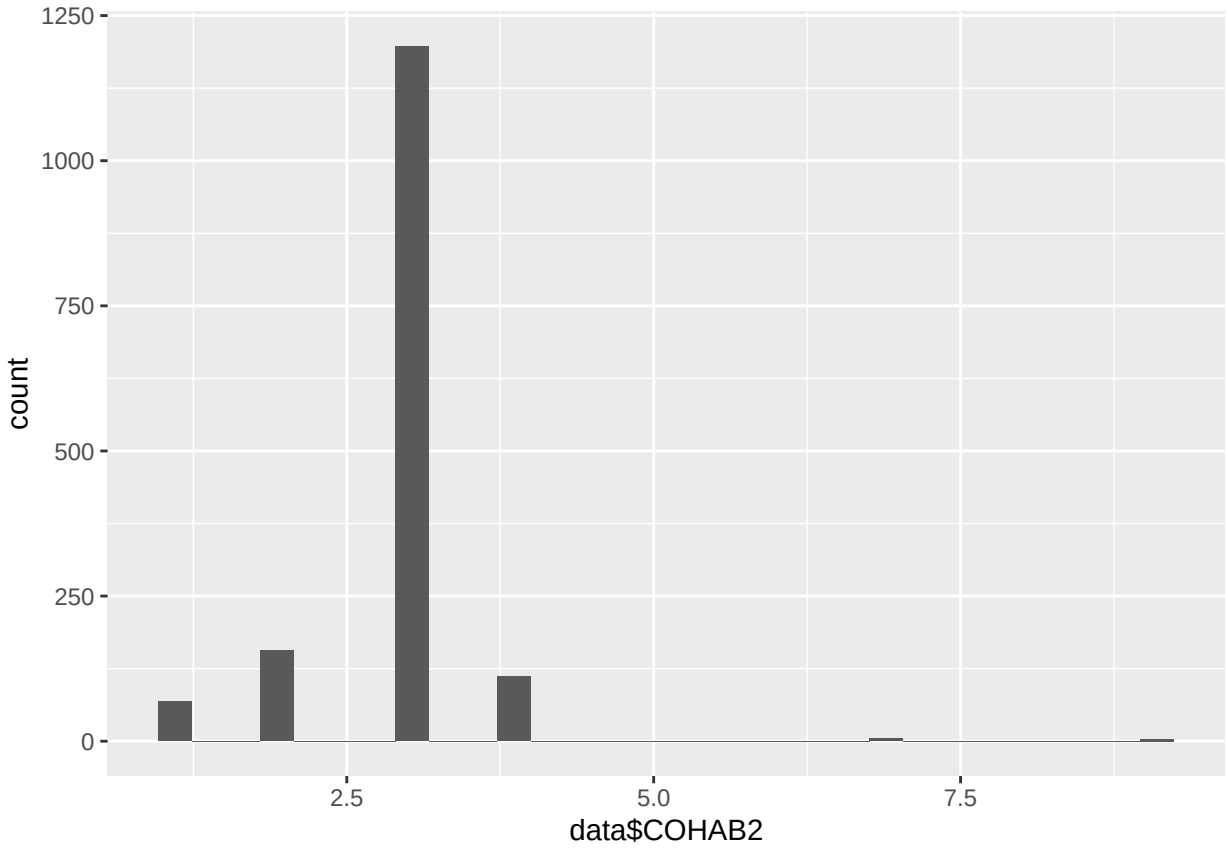
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 68749 rows containing non-finite values (stat_bin).
```




```
ggplot(data, aes(x = data$COHAB2)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 71290 rows containing non-finite values (stat_bin).
```



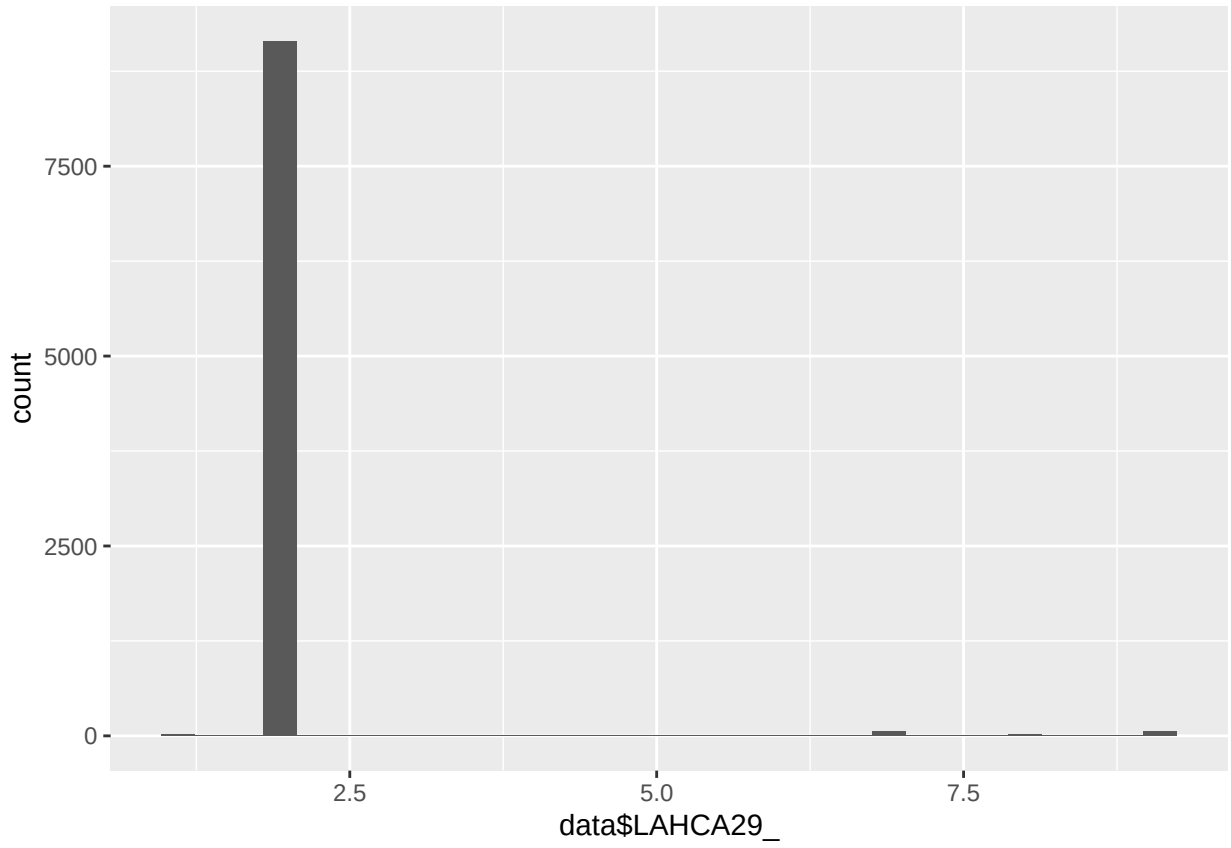
Universe: (AGE GE '018' and AGE not IN ('997','999')) and (PLAADL ='1' or PLAIADL ='1' or PLAWKNOW ='1' or PLAWKLIM ='1' or PLAWALK ='1' or PLAREMEM ='1' or PLIMANY ='1') functional limitation; nervous system; sensory organ condition Nervous system/sensory organ condition causes limitation

1 Mentioned 2 Not mentioned 7 Refused 8 Not ascertained 9 Don't know

```
ggplot(data, aes(x = data$LAHCA29_)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 63543 rows containing non-finite values (stat_bin).
```


7500
5000
2500
0
count
2.5
5.0
7.5
data$LAHCA29_