Exam 1 Solutions

Your name

2024-09-30

Exam 1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1. (5 points) Load the data called dat.nsduh.small.csv. Take a look at the data.

```
dat <- read_csv("dat.nsduh.small.csv")
```

```
## Rows: 171 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (7): mjage, cigage, iralcage, age2, sexatract, speakengl, irsex
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Answer:

2. (5 points) Read the codebook. What do these variables represent: mjage, iralcage, sexatract, speakengl, and irsex?

Answer:

3. (5 points) What type of stat variables are the variables from #2?

```
dat$mjage
```

```
##   [1] 14 11 12 16 14 12 13 20 33 18 13 16 21 12 14 18 18 16 18 20 15 16 15 12 16
##  [26] 19 10 16 17  9 16 16 15 18 13 20 18 16 18 15 14 15 16 17 16 27 22  9 15 18
##  [51] 14 16 20 18 17 15 15 16 18 14  7 17 18 21 19 13 15 18 13 18 11 17 17 15 21
##  [76] 20 14 30 21 13 22 20 16 25 15 15 15 18 14 12 12 35 11 16 21 17 13 21 12 14
## [101]  9 16 16 14 16 19 13 14 17 15 13 17 14 15 27 16 16 12 25 10 14 13 15 13 12
## [126] 13 16 17 13 18 11 14 14 18 11 14 14 15 15 15 14 17 17 16 17 17 32 14 17 14
## [151] 16 15 13 15 16 14 16 15 19 11 16 16 20 13  9 13 16 12 11 17 14
```

```
dat$iralcage
```

```
##   [1] 14  5 12 18 14 18 13 21 16 19 13 16 19 12 17 16 18 12 15 10 15 16 18 14 14
##  [26] 17 14 18 23 11 11 19 12 16 10 18 18 15 18 20 13 12 20 17 15 18 13  9 17 16
##  [51] 18 16 16 17 10 14 19 14 18 15 14 12 18 18 21 14 14 16 12 16  5 15 15 14 15
##  [76] 16 16 18 13 14 21 13 16 21 13 15 15 18 12 12 13 13 11 17 16 16 14 18 12 21
## [101] 18 15 13 13 16 14 14 16 15 17 14 19 14 13 18 16 18 12 18 15 16 13 13 12 12
## [126] 13 12 15 21 17 11  8 14 14 13  8 13 15 16 15 16 12 19 15 16 16 18  7 16 13
## [151] 17 15 14 17 16 12 13 13 17 12 12 14 18 10 14 16 17 15 12 18 13
```

```
dat$sexatract
```

```
##   [1]  1  2  2  1  4  4  1  1  1  1  1  1  1  1  1  1  1  1  1  1  5  1  1  5  2
##  [26]  1  1  1  1 99  1  1  1  2 99  1  1  1  1  2  1  1  1  1  2  1  1  3  1  1
##  [51]  2  1  1  1  1  1  1  1  1  1  1  3  2  1  1  3  1  1  1  1  1  1  1  1  1
##  [76]  1  1  5  1  1  1  1  1  4  1  1  2  1  1  1  1  2  2  1  1  1  6  1  1  1
## [101]  1  1  1  1  1  1  3  1  1  2  3  1  2  1  1  1  1  1  1  1  3  1  1  1  1
## [126]  1  2  3  1  1  3  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## [151]  1  1  1  1  1  1  1  1  1  1  2  1  1  2  1  1  1  1  3  1 99
```

```
dat$speakengl
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 3 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2
## [112] 2 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1
```

```
dat$irsex
```

```
##   [1] 1 2 2 1 1 2 1 1 2 2 1 1 1 2 2 2 1 2 1 1 1 2 1 2 2 2 1 1 2 1 1 1 1 2 1 2 2
##  [38] 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 2 2 1 2 1 1 1 2 2 2
##  [75] 2 2 1 1 1 2 1 2 1 2 2 1 1 2 2 1 1 2 2 2 1 1 1 2 2 1 2 2 2 1 1 2 2 1 2 2 2
## [112] 1 2 2 2 2 2 1 1 2 2 1 1 1 1 1 1 2 2 1 2 1 1 2 1 1 1 1 2 1 2 1 1 1 1 1 2 1
## [149] 1 2 1 2 1 1 1 1 2 1 2 1 2 2 1 2 1 1 1 1 2 2 2
```

Answer:

mjage: Quantitative

iralcage: Quantitative

sexatract: Categorical

speakengl: Categorical

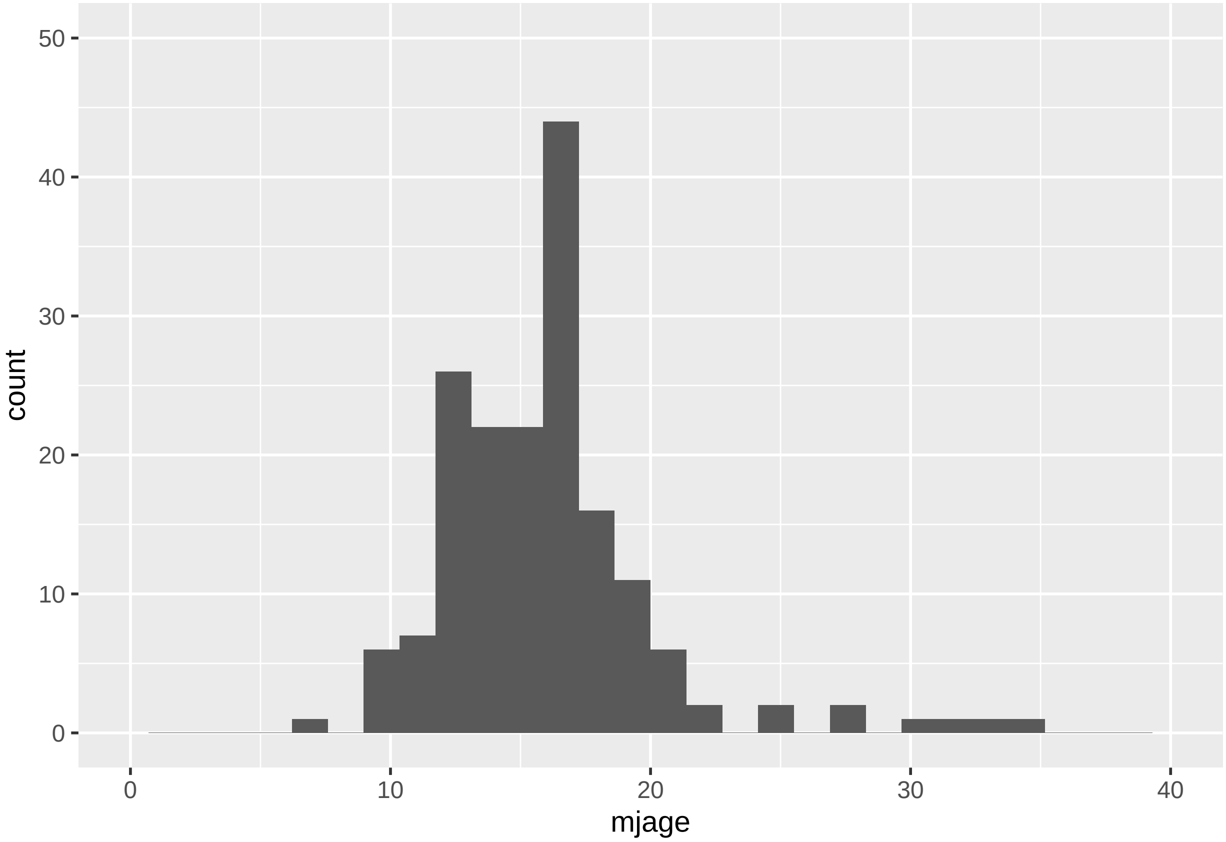
irsex: Categorical

4. (5 points) Do visual EDA for mjage and iralcage, separately. (Hint: To compare the plots, make the axes limits, x and y, be the same for both plots. Also, try out different numbers of bins to see how they change.)

```
#tidyverse version
dat %>% ggplot(aes(mjage)) + geom_histogram() + xlim(0,40) + ylim(0,50)
```

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

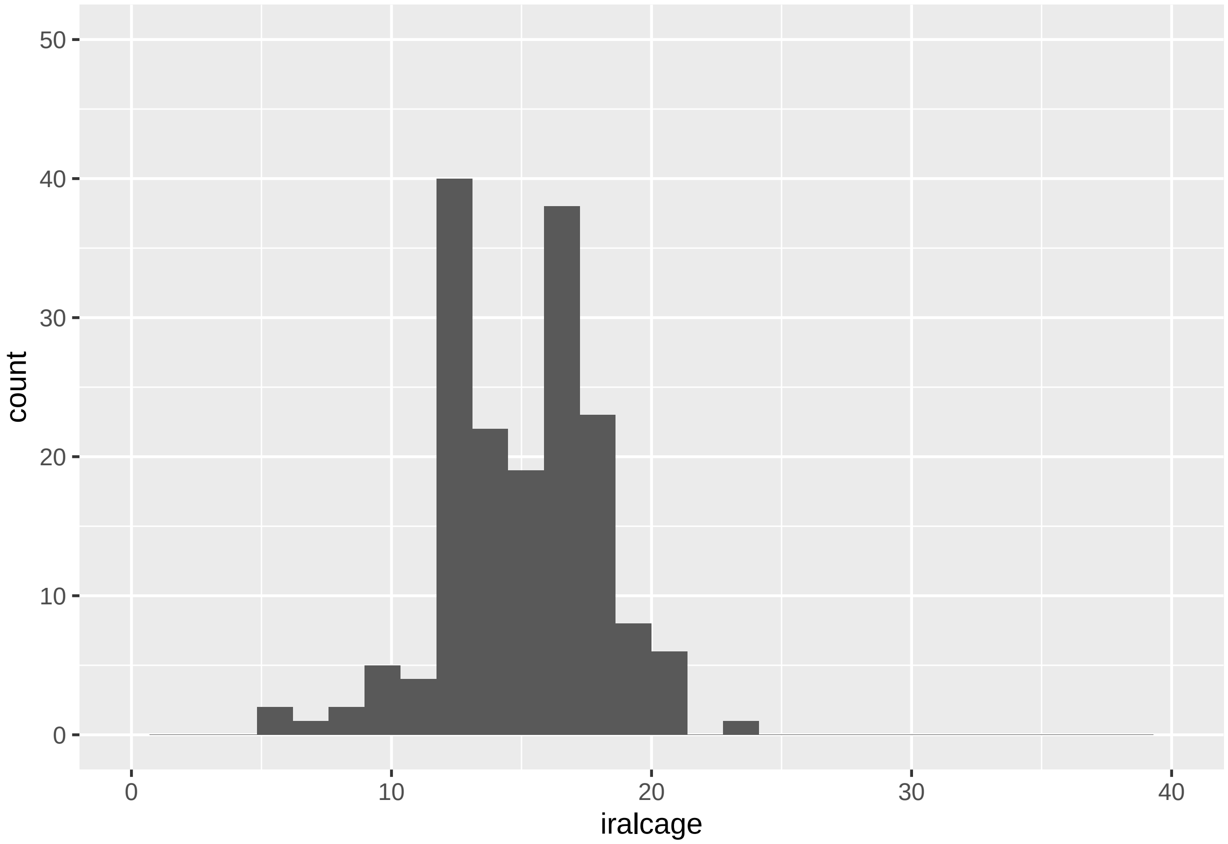
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



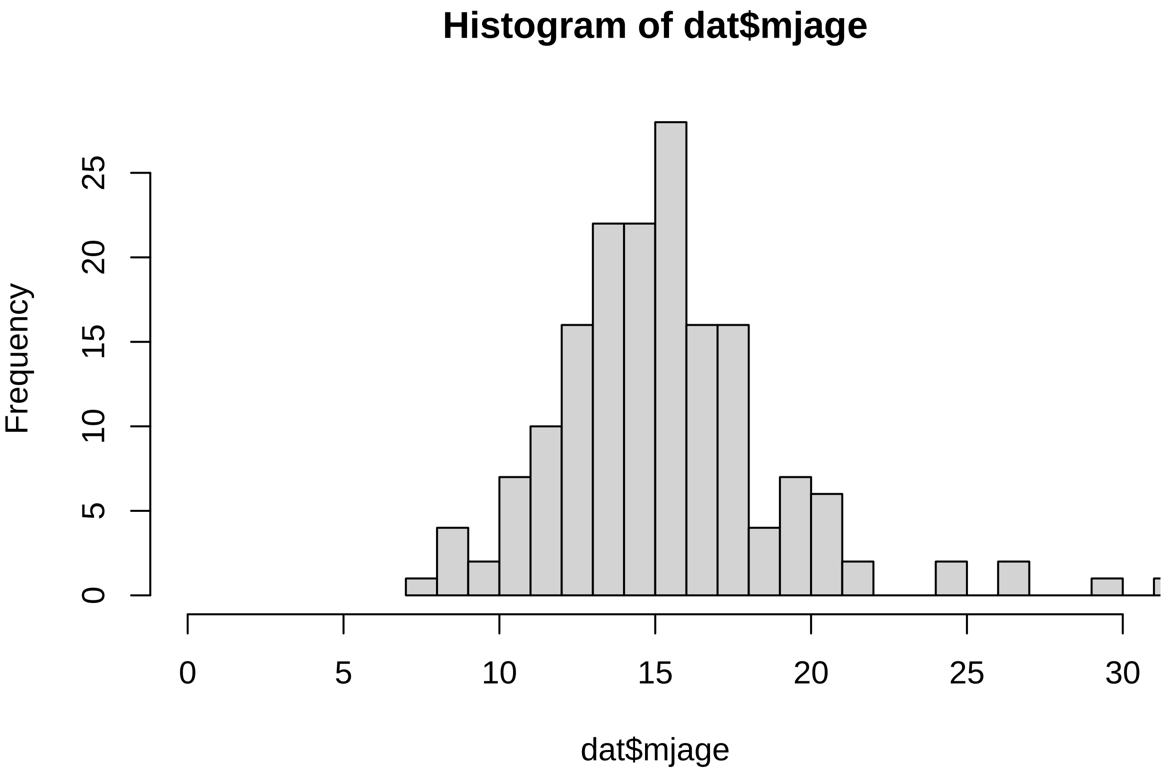
```
dat %>% ggplot(aes(iralcage)) + geom_histogram() + xlim(0,40) + ylim(0,50)
```

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

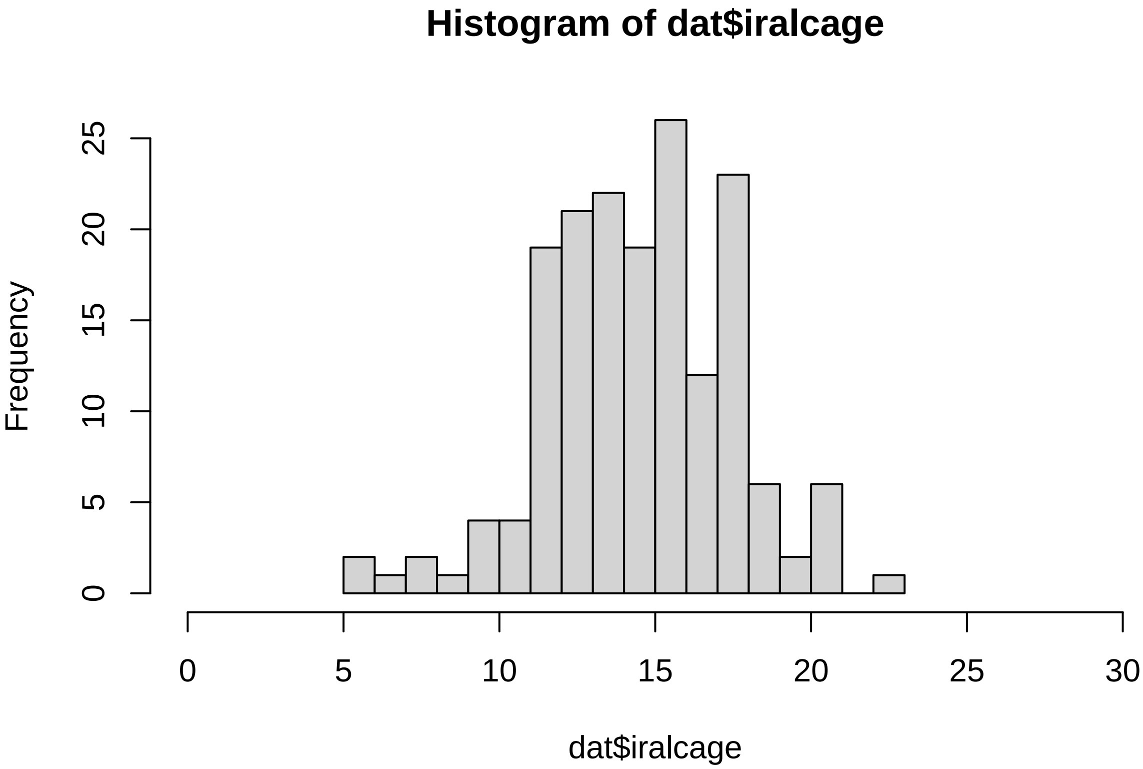
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



```
# base R version
hist(dat$mjage, breaks = 20, xlim=c(0,30))
```



```r
hist(dat$iralcage, breaks = 20, xlim=c(0,30))
```



Answer:

mjage: Unimodal, symmetric, longer right tail, no clear outliers.

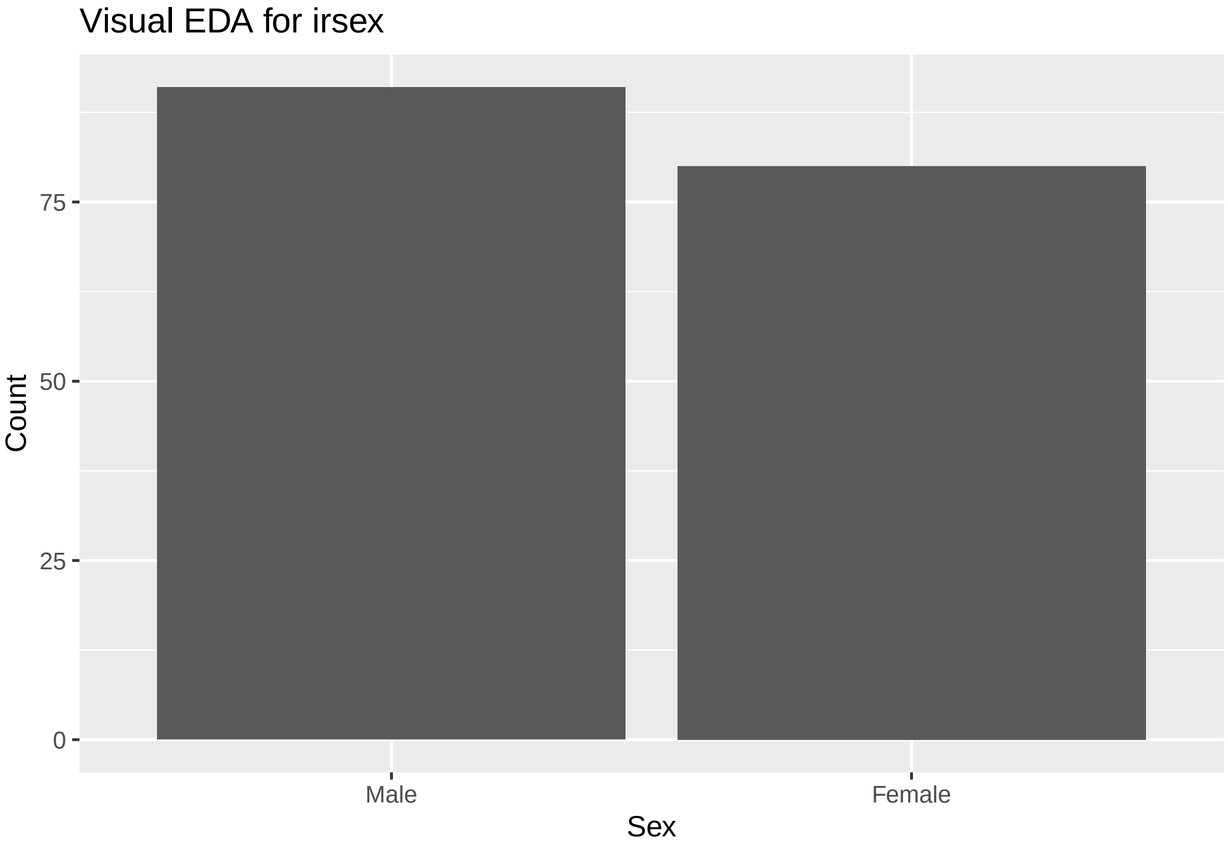
iralcage: Bimodal, symmetric, longer left tail, no clear outliers.

5. (5 points) Compare the two plots. How are they different and how are they similar? What theory do you have to explain these differences?

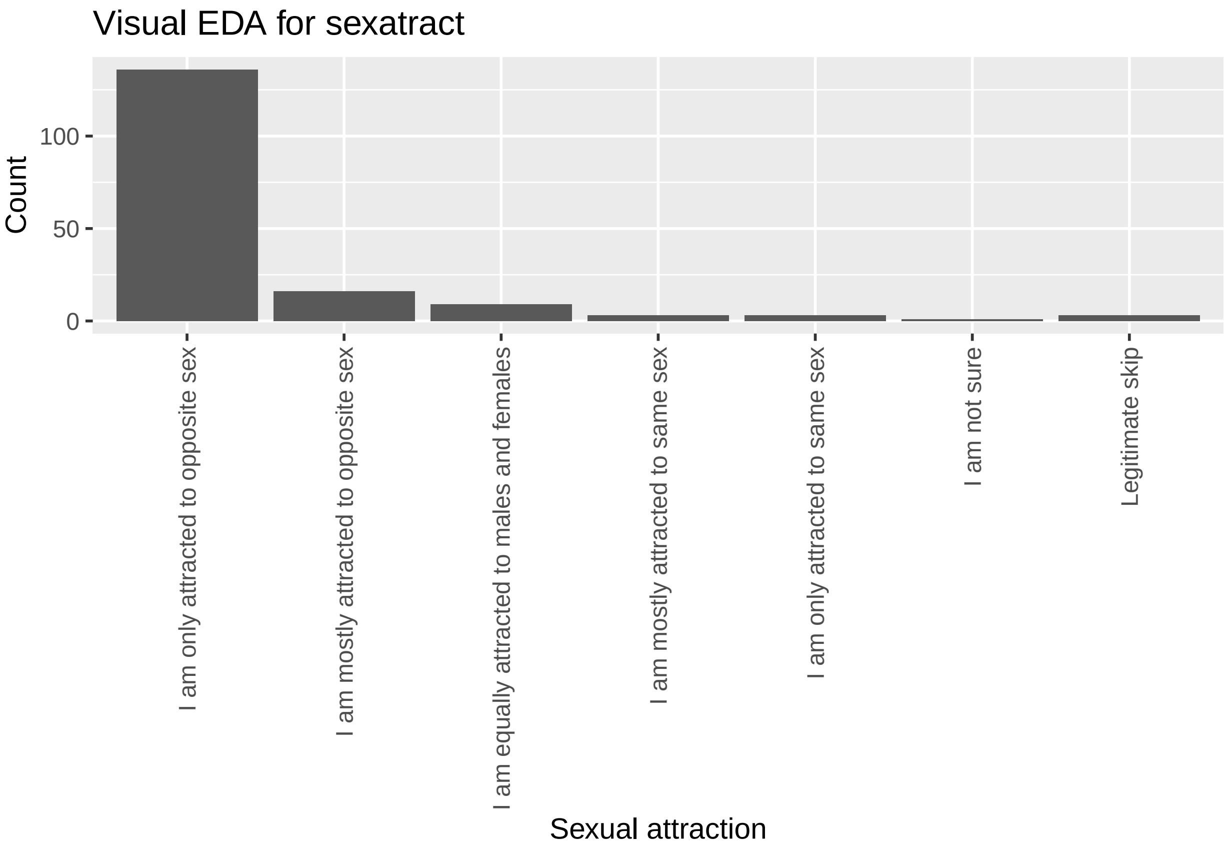
Answer: They're similar, but iralcage is bimodal (perhaps some people try alcohol for the first time in high school and some try it in college) and has a longer left tail, meaning that people start using alcohol when they are younger, rather than older. For marijuana, there is a longer right tail, which means that some people start using when they are older.

6. (5 points) Do visual EDA for irsex and sexatract. What do you learn from this?

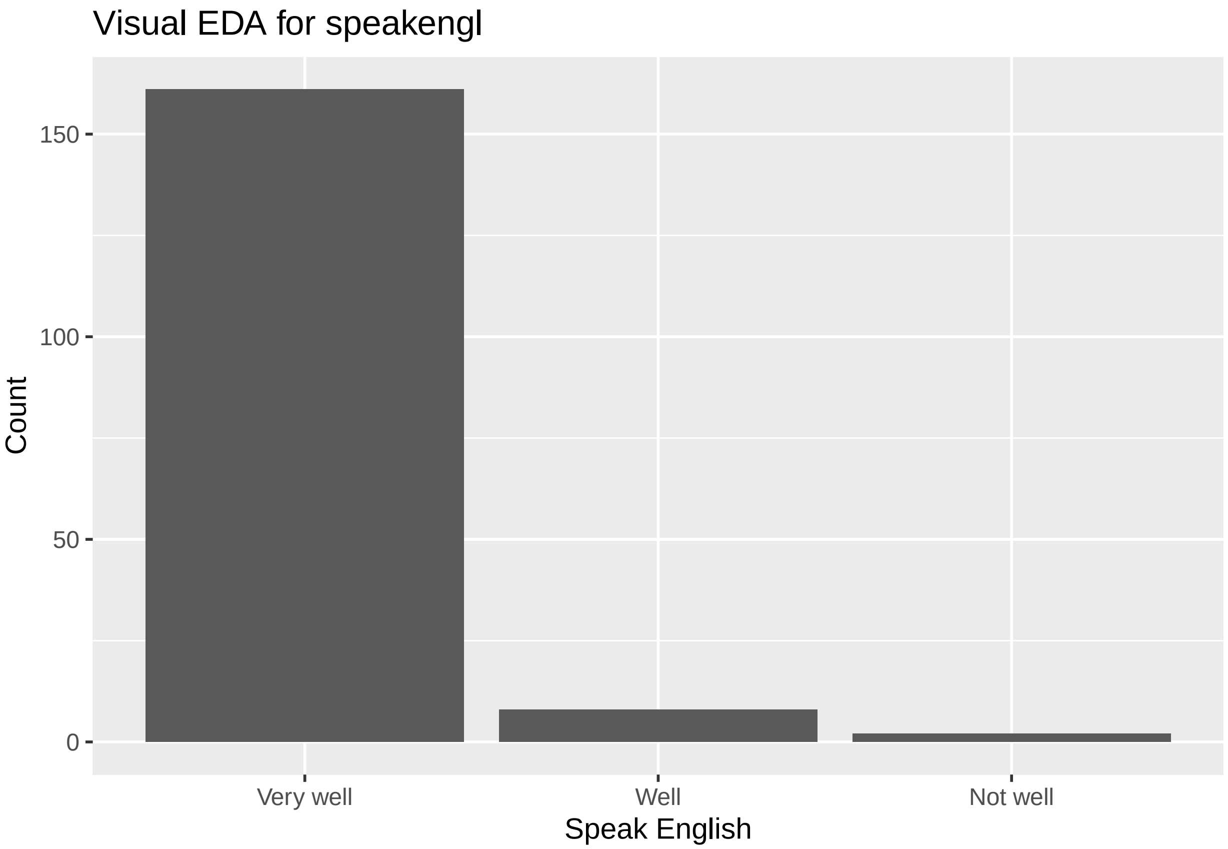
```r
# this code makes the levels of the factors take on the values from the codebook:
dat <- dat %>%
  mutate_at(c('irsex', 'sexatract', 'speakengl'), as.factor) %>%
  mutate(irsex = recode(irsex,          "1"="Male", "2"="Female"),
         sexatract = recode(sexatract, "1"="I am only attracted to opposite sex",
                                       "2"="I am mostly attracted to opposite sex",
                                       "3"="I am equally attracted to males and females",
                                       "4"="I am mostly attracted to same sex",
                                       "5"="I am only attracted to same sex",
                                       "6"="I am not sure",
                                       "99"="Legitimate skip"),
         speakengl = recode(speakengl, "1"="Very well",
                                       "2"="Well",
                                       "3"="Not well",
                                       "4"="Not at all"))

dat %>%
  ggplot(aes(irsex)) +
  geom_bar() +
  xlab("Sex") +
  ylab("Count") +
  ggtitle("Visual EDA for irsex")
```



```
dat %>%
  ggplot(aes(sexatract)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1)) +
  xlab("Sexual attraction") +
  ylab("Count") +
  ggtitle("Visual EDA for sexatract")
```



```
dat %>% ggplot(aes(speakengl)) + geom_bar() + xlab("Speak English") + ylab("Count") + ggtitle("Visual E
```



Answer: There are more men than women in this sample, most people report only being attracted to the opposite sex, most people speak English very well.

7. (5 points) Do quantitative EDA for irsex, sexatract, and speakengl. What do you learn from this that you didn't know from the visual EDA?

```
dat %>%
  count(irsex)%>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 2 x 3
##   irsex      n  prop
##   <fct>  <int> <dbl>
## 1 Male      91 0.532
## 2 Female    80 0.468
```

```
dat %>%
  count(sexatract)%>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 7 x 3
##   sexatract                                       n    prop
##   <fct>                                       <int>   <dbl>
## 1 I am only attracted to opposite sex           136 0.795
## 2 I am mostly attracted to opposite sex          16 0.0936
## 3 I am equally attracted to males and females     9 0.0526
## 4 I am mostly attracted to same sex               3 0.0175
## 5 I am only attracted to same sex                       3 0.0175
## 6 I am not sure                                         1 0.00585
## 7 Legitimate skip                                       3 0.0175
```

```
dat %>%
  count(speakengl)%>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 3 x 3
##   speakengl     n   prop
##   <fct>     <int>  <dbl>
## 1 Very well   161 0.942
## 2 Well          8 0.0468
## 3 Not well      2 0.0117
```

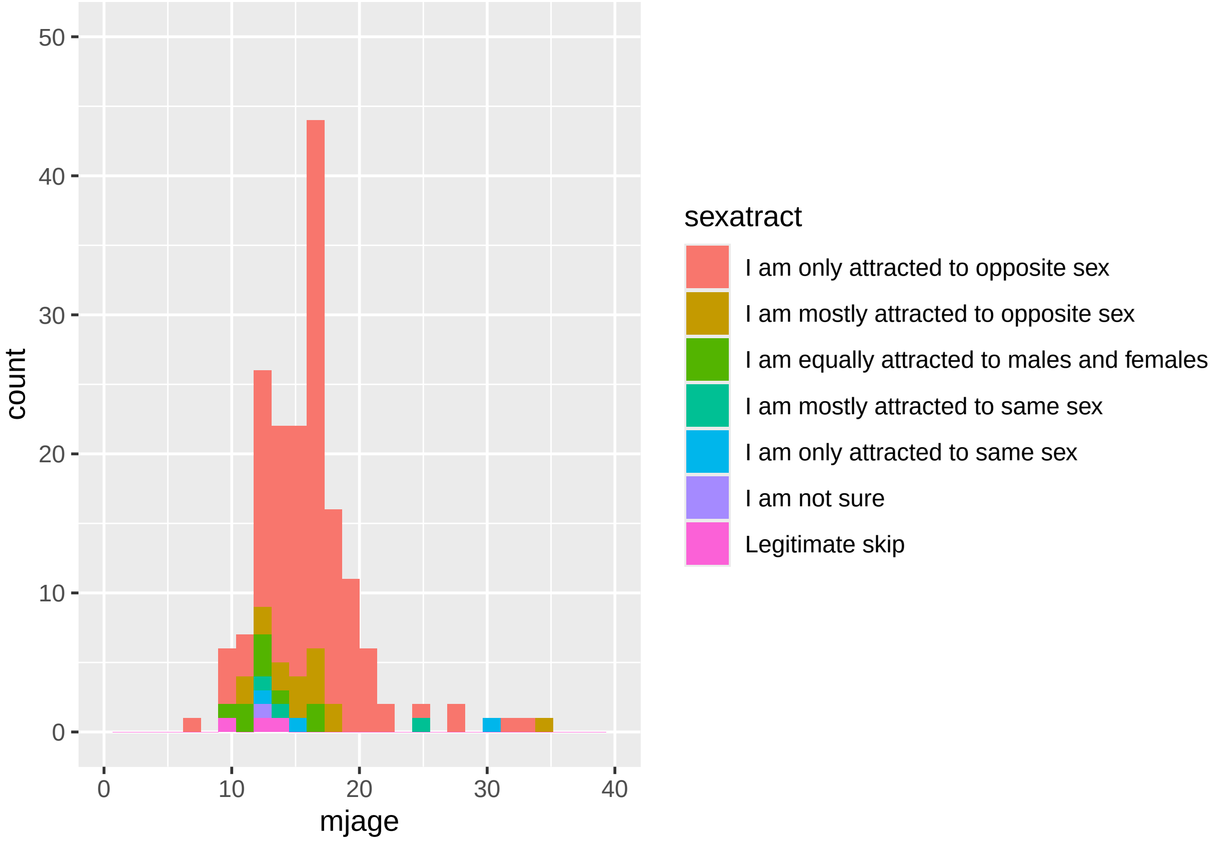
Answer: The small numbers of the lower four categories.

8. (5 points) Copy your line from the visual EDA for mjage. What happens when you use this code "aes(x=mjage, fill=sexatract)" for the aesethetics in ggplot? Do you find out something interesting?

```
dat %>% ggplot(aes(x=mjage, fill=sexatract)) + geom_histogram() + xlim(0,40) + ylim(0,50)
```

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

```
## Warning: Removed 14 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



Answer: It looks like the individuals who are not just attracted to the opposite sex try marijuana at younger ages compared to those who are attracted only to the opposite sex.