Exploratory Data Analysis EDAwR-Slides2.pdf

Read in the data

Read full file *once* then sample, save, comment out.

```
# df <- read.csv("~/Downloads/nat2021us.csv")
# set.seed(504)
# df_samp <- df[sample(nrow(df), 1000),]
# write.csv(df_samp, "sample1000.csv", row.names = FALSE)
# dat <- df[, c("mager", "fagecomb", "meduc", "dmar")]
# write.csv(dat, "births2021.csv", row.names = FALSE)
dat <- read.csv("births2021.csv")
```

Load libraries, etc.

```
library(tidyverse)
options(scipen = 999) # no scientific notation
theme_set(theme_bw(16)) # set ggplot2 theme
```

The data

```
1 head(dat)
```

```
  mager fagecomb meduc dmar
1    22       23     3    1
2    31       27     4    2
3    29       29     4    2
4    39       40     6    2
5    20       28     3    2
6    29       39     3    1
```

```
1 dim(dat)
```

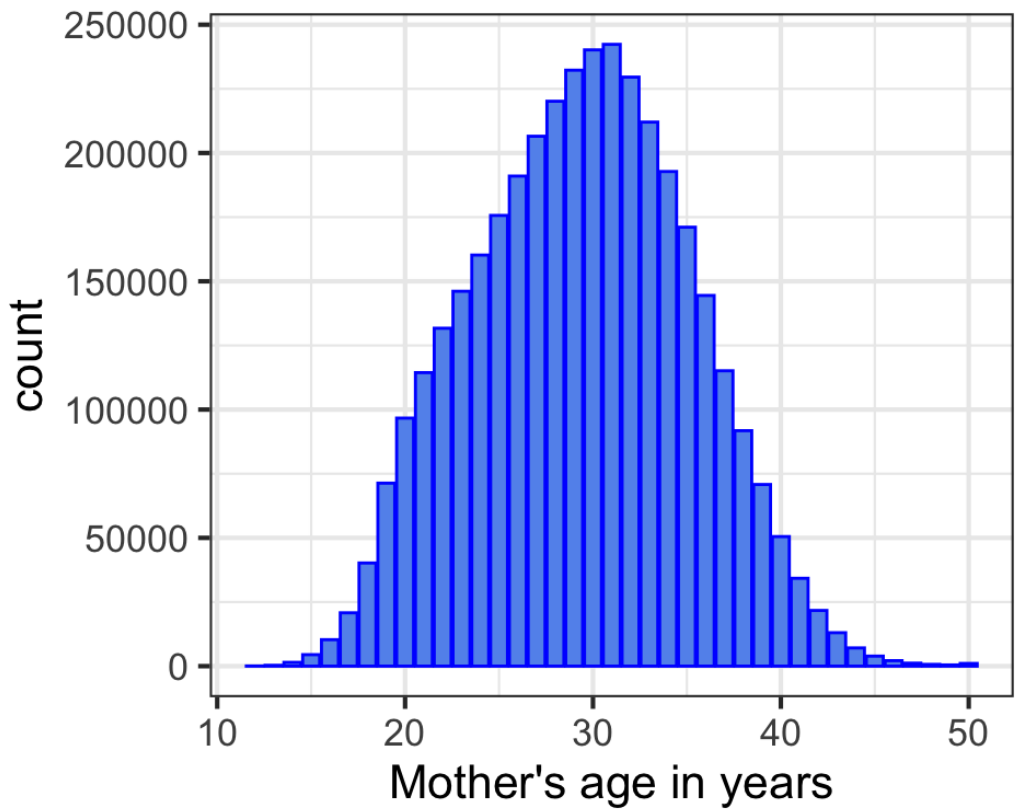
```
[1] 3669928       4
```

format: *cases* or *unbinned*

(no count or frequency column)

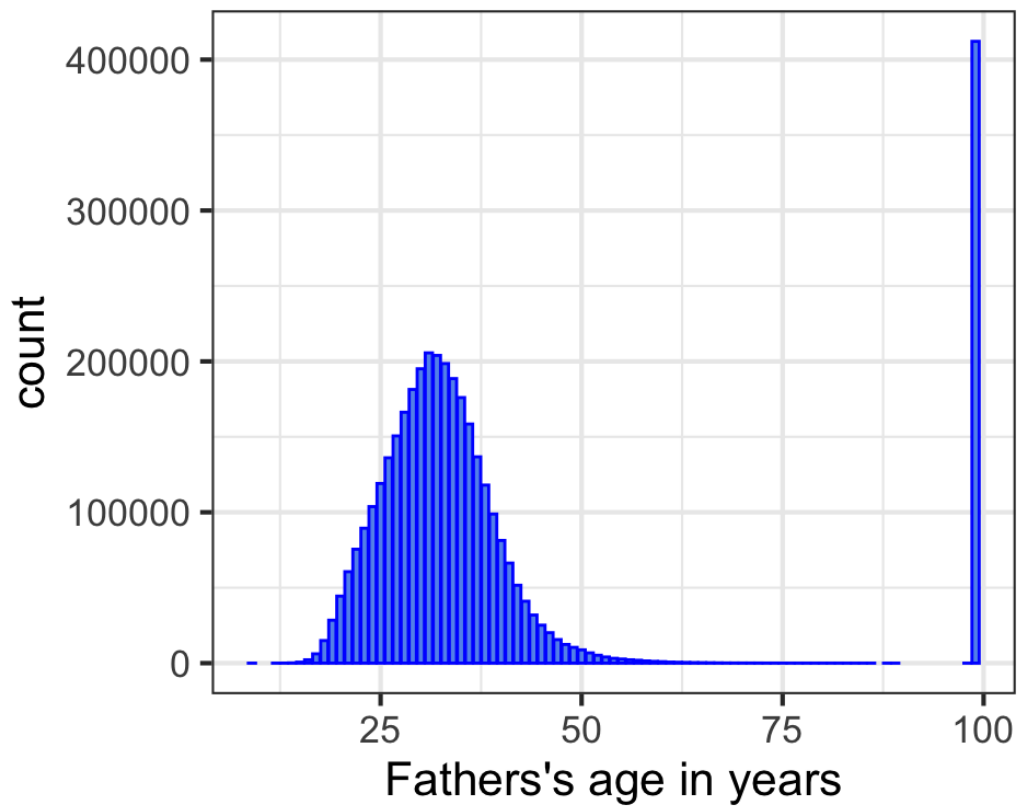
Mother’s age

```
ggplot(dat, aes(mager)) +
  geom_bar(color = "blue", fill = "cornflowerblue") +
  xlab("Mother's age in years")
```



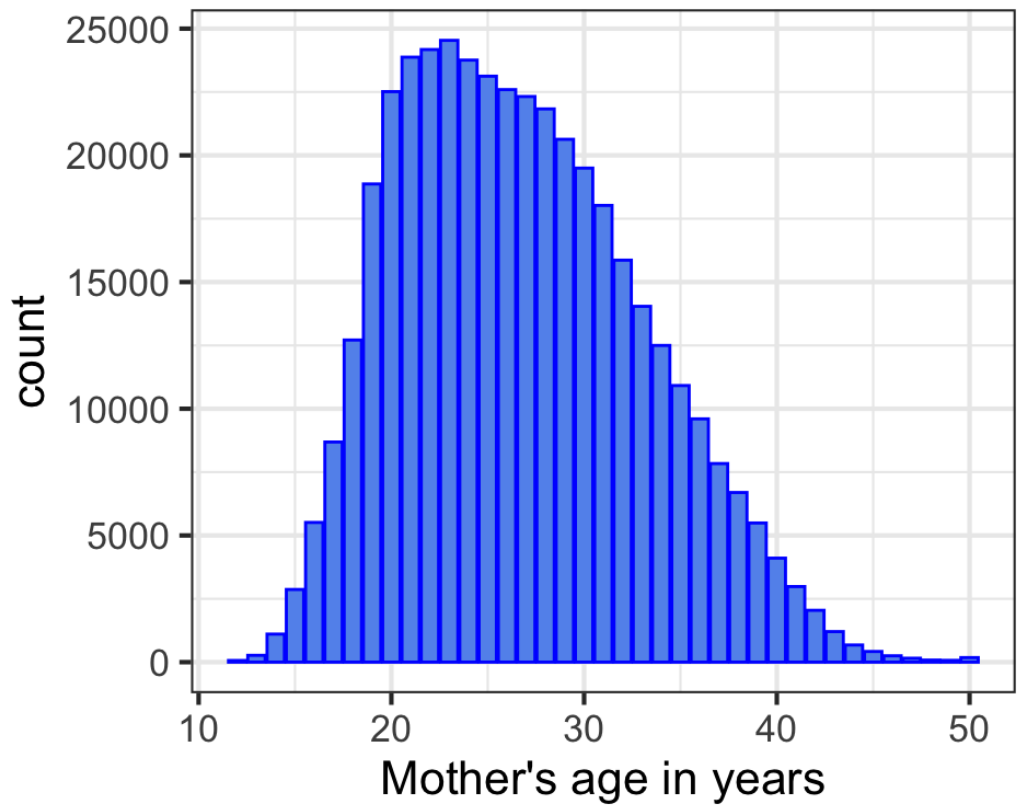
Father's age

```
ggplot(dat, aes(fagecomb)) +
  geom_bar(color = "blue", fill = "cornflowerblue") +
  xlab("Fathers's age in years")
```



Mother's age (father's age is missing)

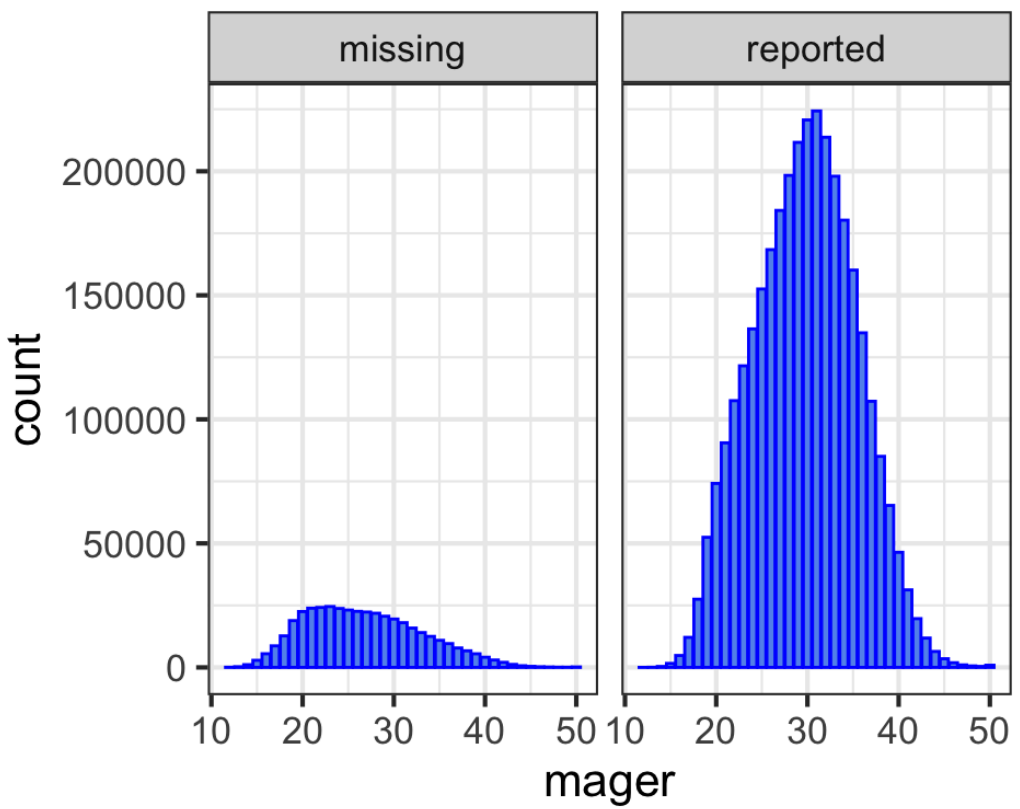
```
dat |> filter(fagecomb == 99) |>
ggplot(aes(mager)) +
  geom_bar(color = "blue", fill = "cornflowerblue") +
  xlab("Mother's age in years")
```



Mother's age by father's age reporting

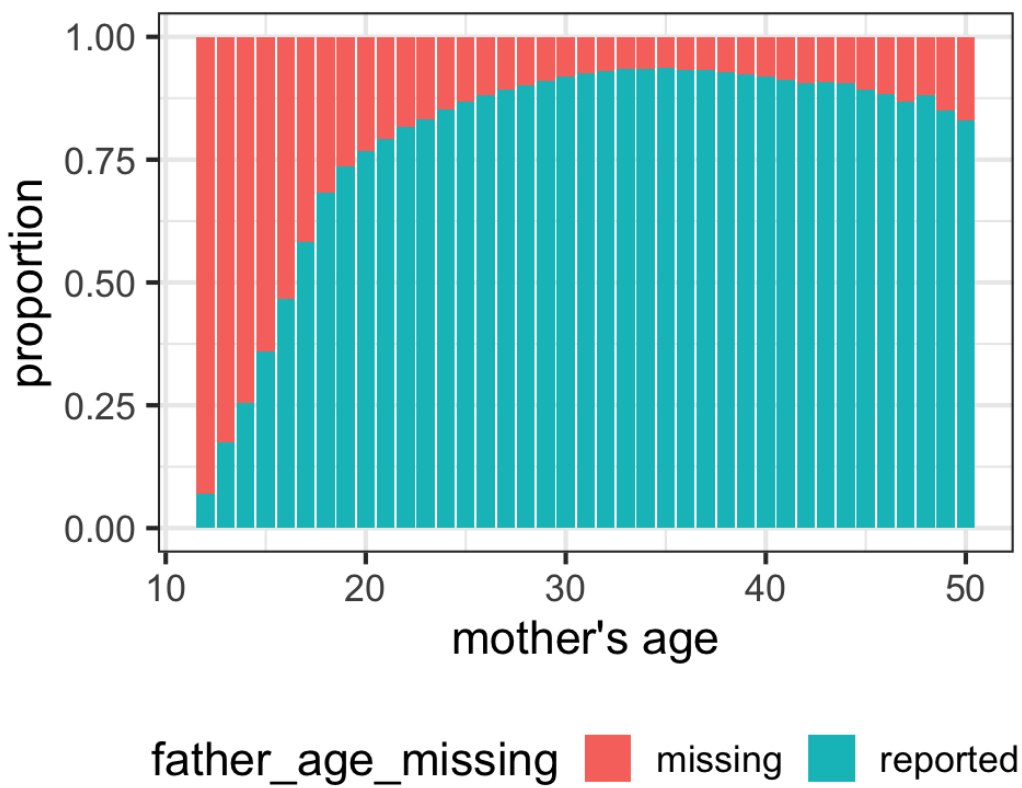
Create a new variable to indicate missing status

```
dat$father_age_missing <- ifelse(dat$fagecomb == 99, "missing", "reported")
ggplot(dat, aes(mager)) + geom_bar(color = "blue", fill = "cornflowerblue") +
  facet_wrap(~father_age_missing)
```



Mother’s age by father’s age reporting

```
dat |>
  ggplot(aes(mager, fill = father_age_missing)) +
  geom_bar(position = "fill") +
  labs(x = "mother's age", y = "proportion") +
  theme(legend.position = "bottom")
```



Father's age vs. mother's age

Bin data first. (Not practical to make a scatterplot with almost 4 million data points.)

```
df_count <- dat |> group_by(mager, fagecomb) |> count()
head(df_count)
```

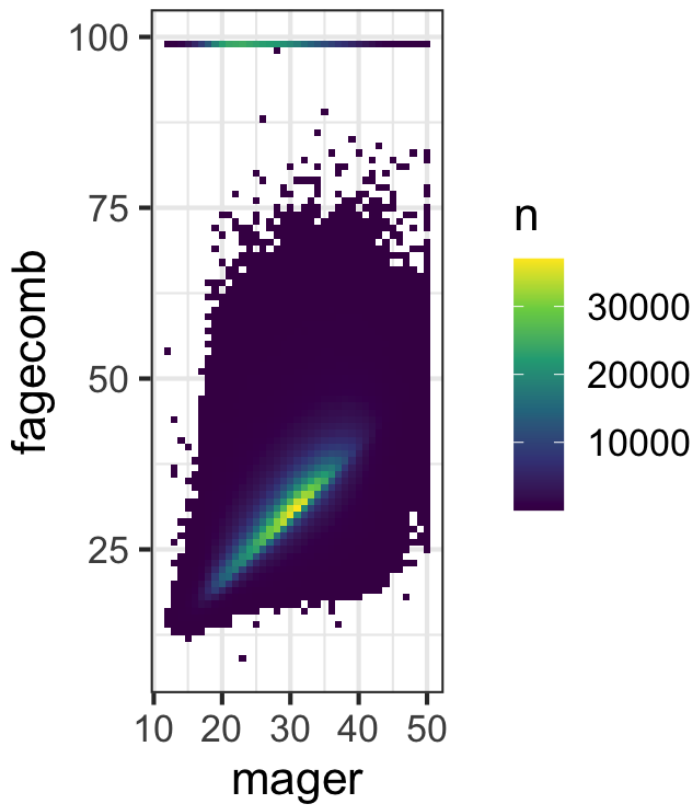
```
# A tibble: 6 × 3
# Groups:   mager, fagecomb [6]
  mager fagecomb     n
  <int>    <int> <int>
1    12       14     1
2    12       15     1
3    12       16     2
4    12       54     1
5    12       99    66
6    13       13     8
```

```
dim(df_count)
```

```
[1] 1999    3
```

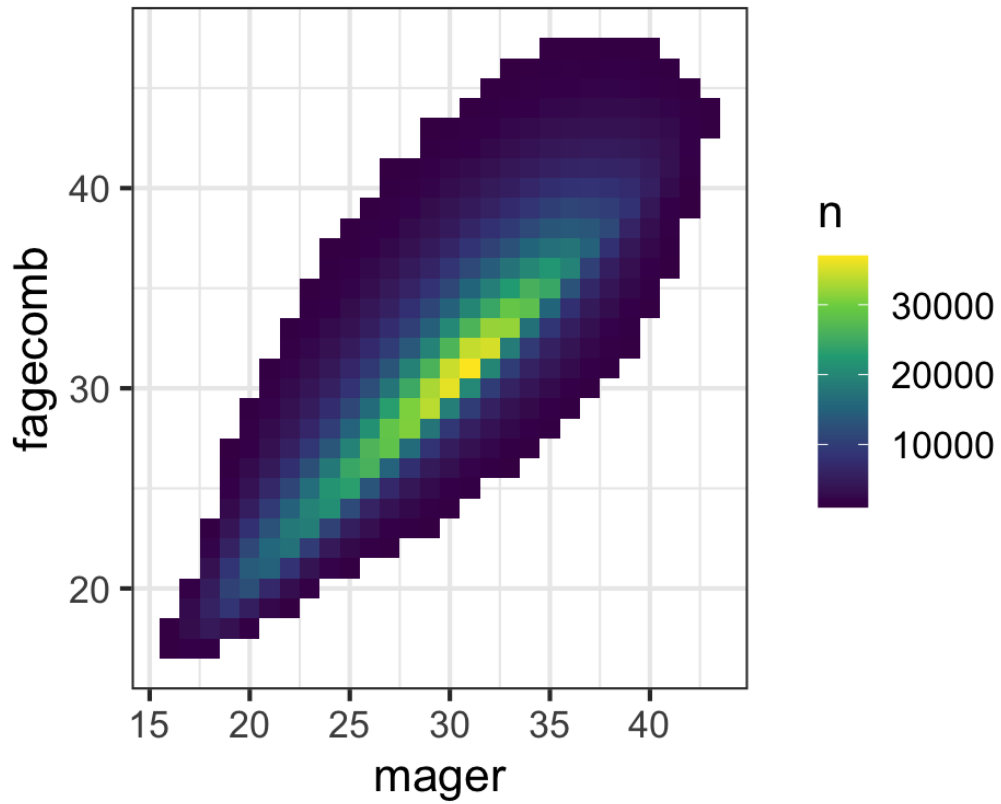
Father's age vs. mother's age

```
ggplot(df_count, aes(mager, fagecomb, fill = n)) +
  geom_tile() +
  scale_fill_viridis_c() +
  coord_fixed()
```



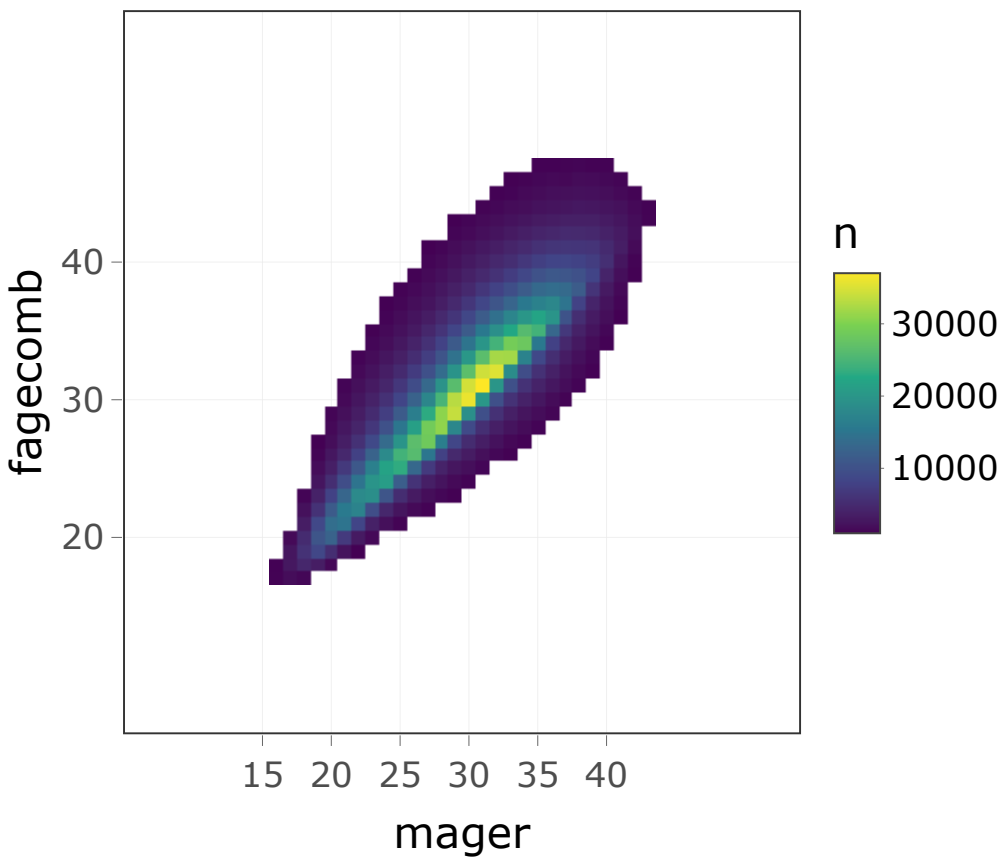
Father's age vs. mother's age

```
g <- df_count |> filter(n > 1000, fagecomb != 99) |>
  ggplot(aes(mager, fagecomb, fill = n)) +
  geom_tile() +
  scale_fill_viridis_c() +
  coord_fixed()
g
```


fagecomb
40
30
20
15
20
25
30
35
40
mager
n
30000
20000
10000

Father’s age vs. mother’s age

```
1 plotly::ggplotly(g)
```



Most popular ages

```r
df_count |> group_by(mager) |> summarize(count = sum(n)) |>
  slice_max(count, n = 4)
```

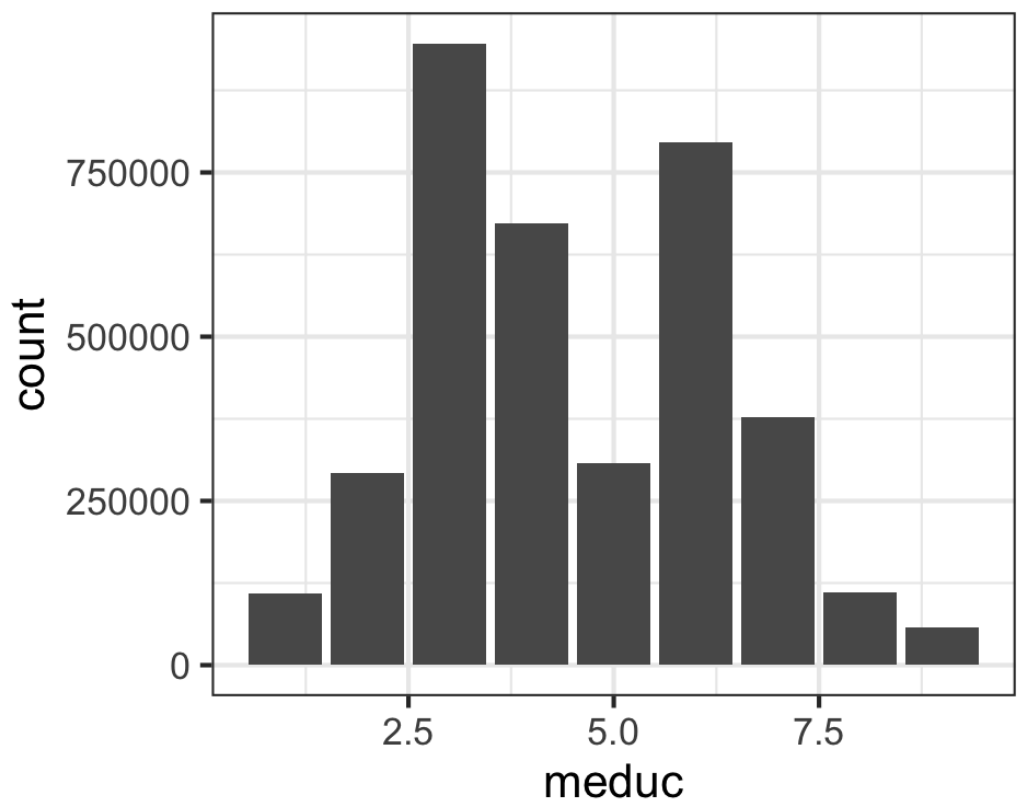
```
# A tibble: 4 × 2
  mager  count
  <int>  <int>
1    31 242336
2    30 240183
3    29 232272
4    32 229576
```

```r
df_count |> group_by(fagecomb) |> summarize(count = sum(n)) |>
  slice_max(count, n = 4)
```

```
# A tibble: 4 × 2
  fagecomb  count
     <int>  <int>
1       99 412109
2       31 205647
3       32 203960
4       33 198552
```

Mother's education

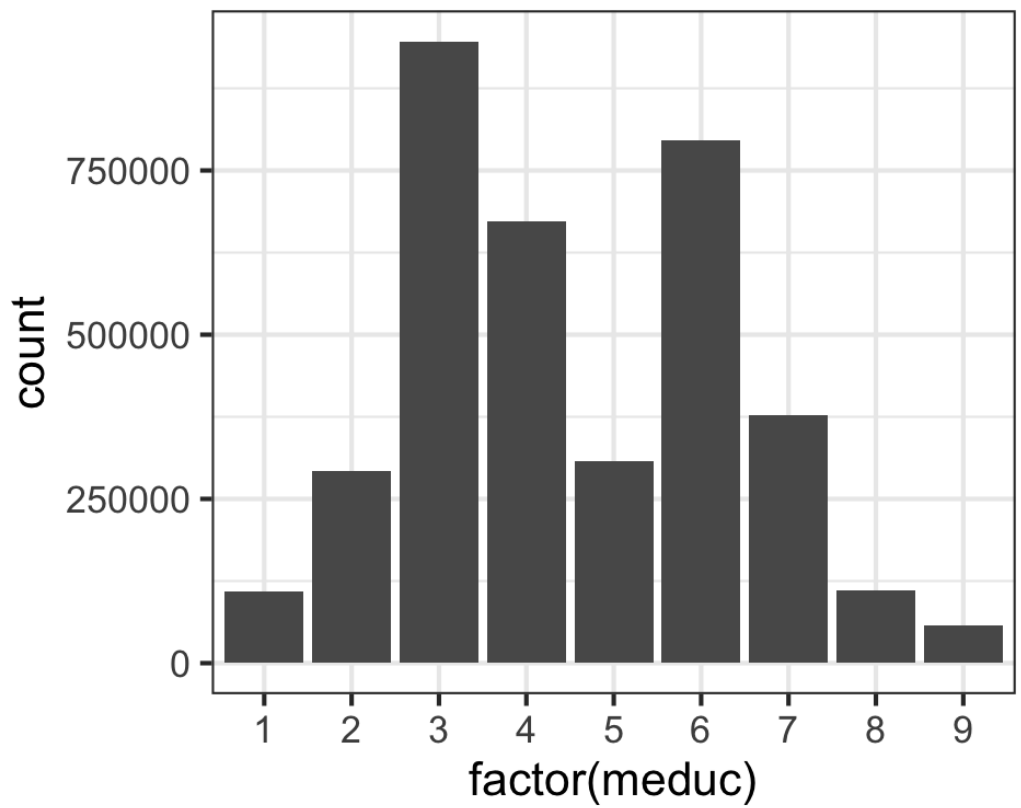
```
1  ggplot(dat, aes(meduc)) + geom_bar()
```



(x-axis is incorrect)

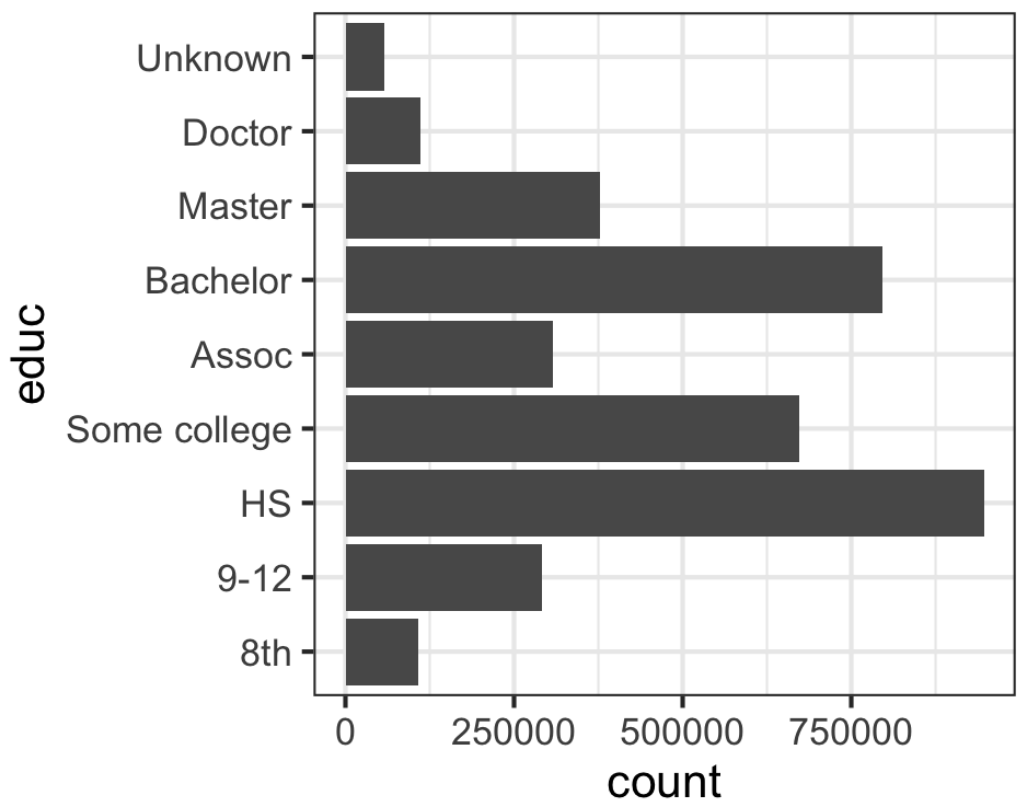
Mother's education

```
1  ggplot(dat, aes(factor(meduc))) + geom_bar()
```



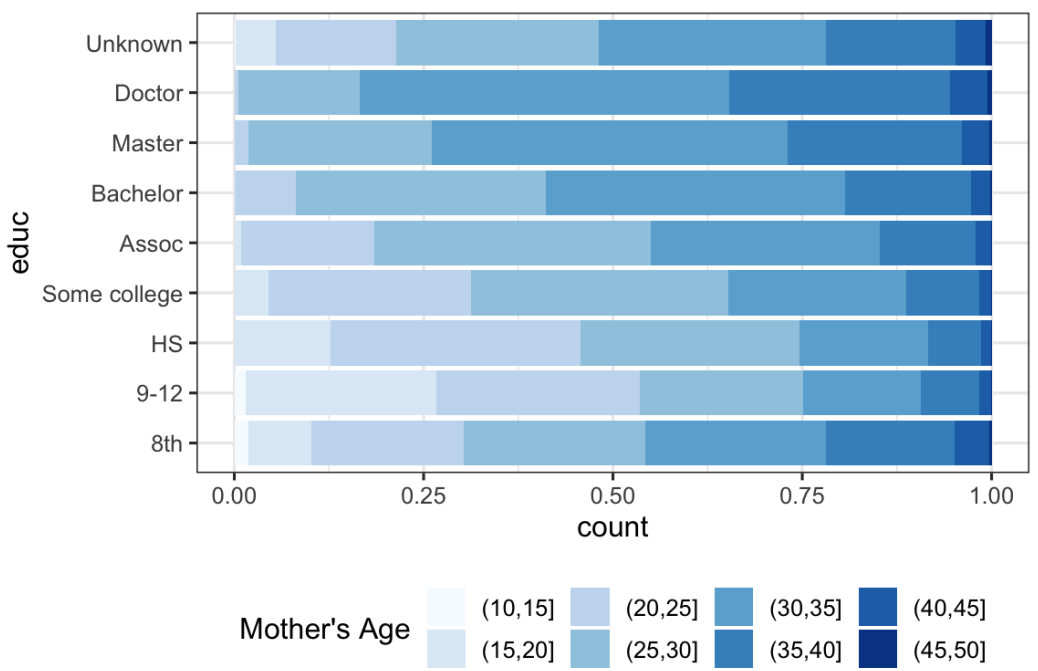
Mother's education

```
dat$educ <- fct_recode(factor(dat$meduc), "8th" = "1", "9-12" = "2",
                       "HS" = "3", "Some college" = "4", "Assoc"= "5",
                       "Bachelor" = "6", "Master" = "7", "Doctor" = "8",
                       "Unknown" = "9")
ggplot(dat, aes(educ)) + geom_bar() + coord_flip()
```



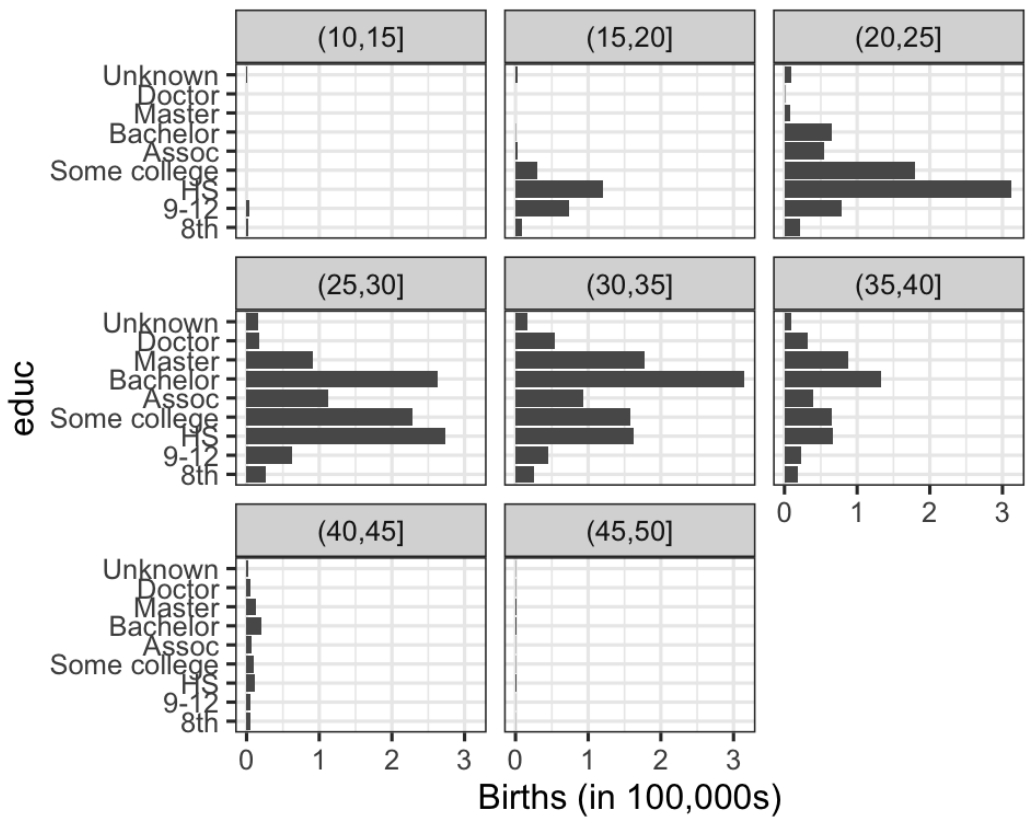
Mother's age and education

```
dat$m_age <- cut(dat$mager, breaks = seq(10, 50, 5))
ggplot(dat, aes(educ, fill = fct_rev(m_age))) +
  geom_bar(position = "fill") +
  coord_flip() +
  scale_fill_brewer(direction = -1, breaks = levels(dat$m_age)) +
  guides(fill=guide_legend(title="Mother's Age")) +
  theme_bw(12) +
  theme(legend.position = "bottom")
```



Mother's age and education

```
ggplot(dat, aes(educ)) +
  geom_bar() + facet_wrap(~m_age) + coord_flip() +
  scale_y_continuous(labels = ~.x/100000) +
  ylab("Births (in 100,000s)") + theme_bw(12)
```



Mother's age and education

```
ggplot(dat, aes(m_age)) + geom_bar() + facet_wrap(~educ) +
  coord_flip() + scale_y_continuous(labels = ~.x/100000) +
  ylab("Births (in 100,000s)") + theme_bw(12)
```