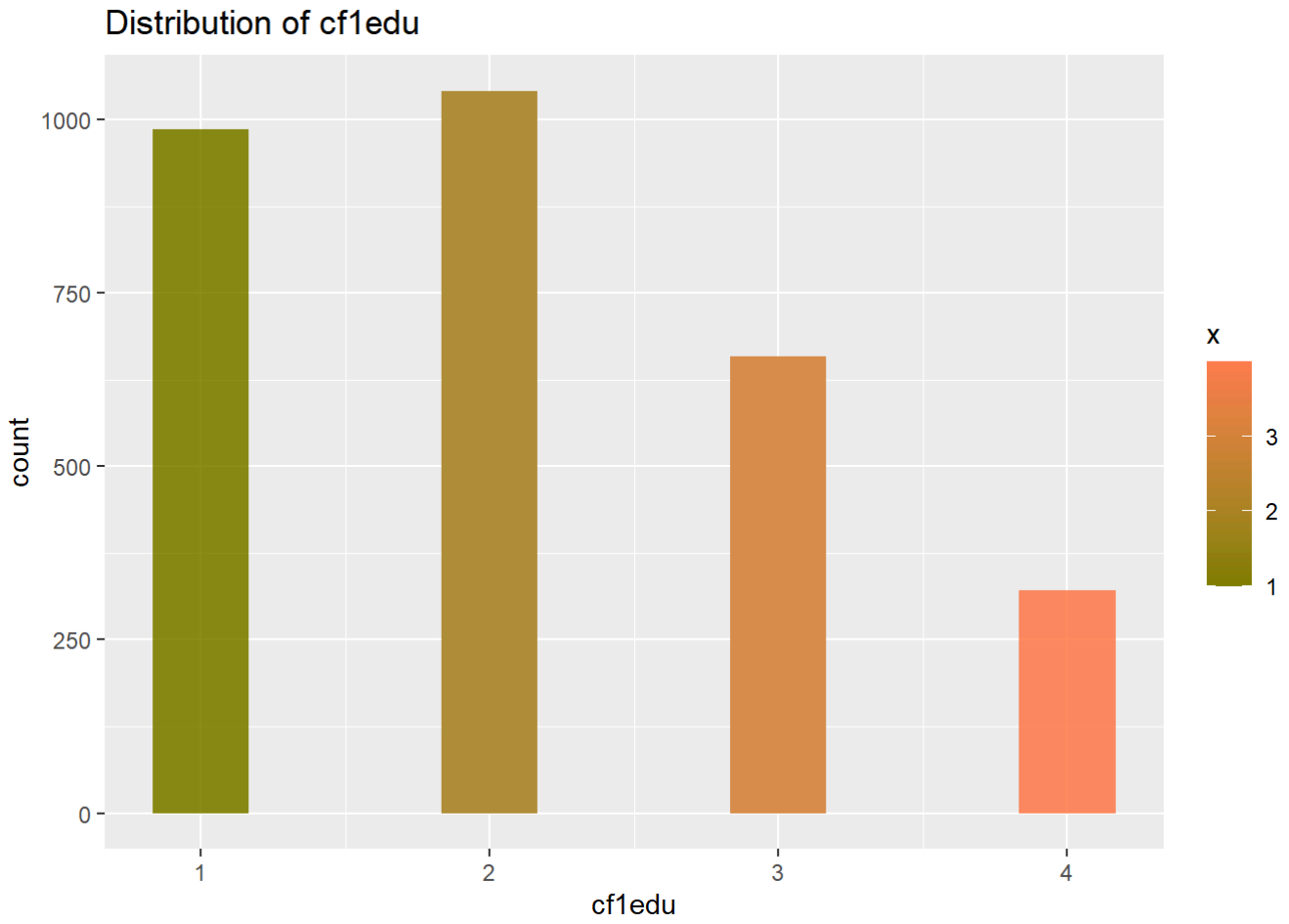
EDA

```
# Load the haven package
library(haven)

# Replace "your_file.dta" with the path to your Stata dataset
data <- read_dta("D:/BrownUnivercity/DATA2020/Data2020-Final-Project/data/ff_data_x_preprocesse
d.dta")

print(data)
```

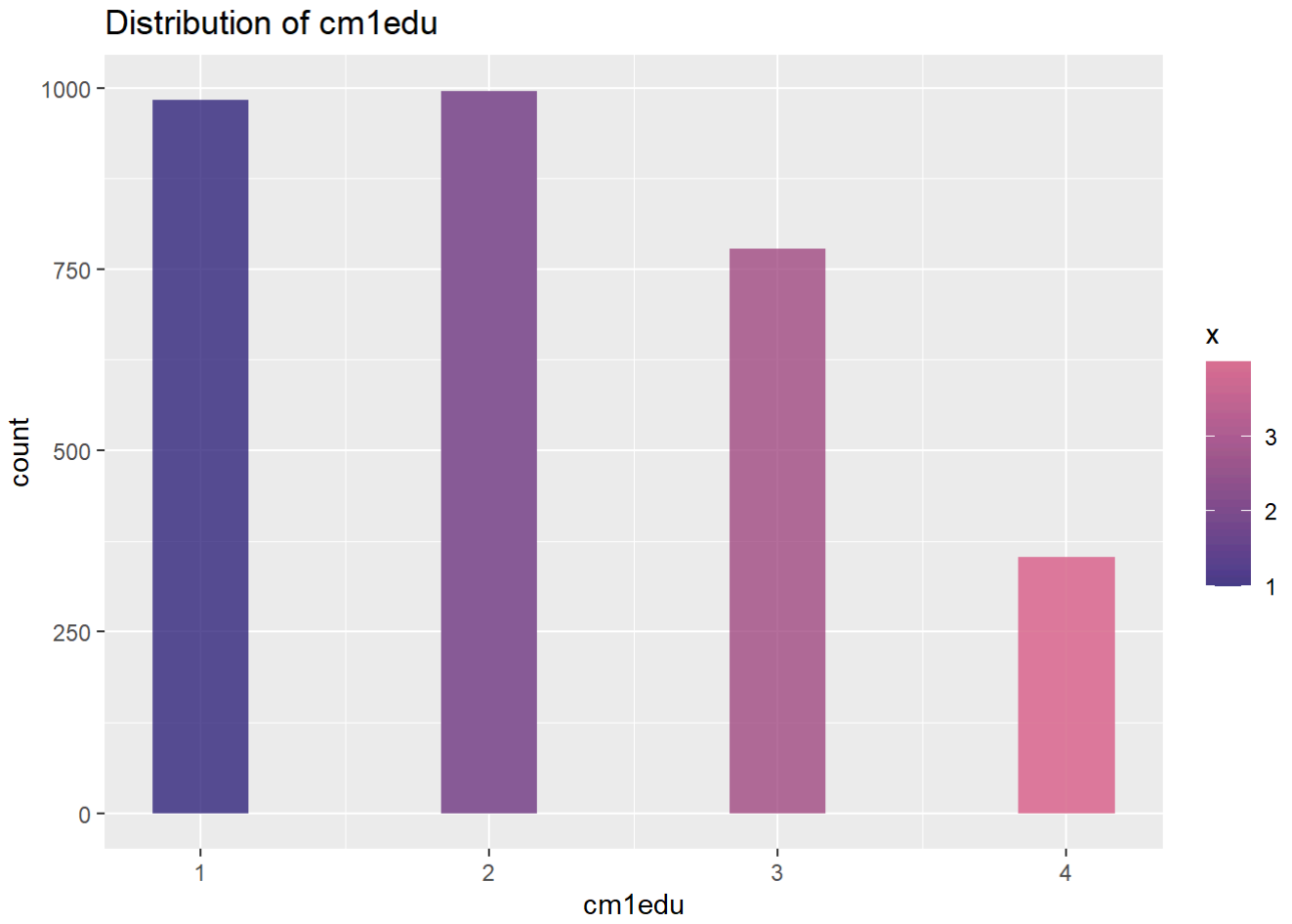
```
## # A tibble: 3,113 × 32
##    index cfledu cmledu cflhhinc  flj8 mli2b f2b13 f2b32 f2d1a  m2d2 m2b43a m2b10
##    <dbl>  <dbl>  <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
##  1     0      3      3    22500 32000    10    -6    -6    -5     1     -6    -6
##  2     1      1      1       -9    -9   260    -6     1     2     1     -6     3
##  3     2      3      3    62500 20000     6    -6     2     2     1     -6    -6
##  4     3      2      2    30000  1800   400    -6     1     1     1     -6    -6
##  5     5      2      2       -9    -9     6    -9    -9    -9    -6     -6     1
##  6     7      2      1    21063     6     5    -6     1     1     1     -6    -6
##  7     8      1      3    30000    10     7    -6    -6    -5     1     -6    -6
##  8     9      2      3       -9    -9    -1    -9    -9    -9    -9     -9    -9
##  9    10      2      3    22500   300   150    -9    -9    -9     1     -6     1
## 10    11      2      3     3750   200     6    -6     3     1     2     -6    -6
## # ℹ 3,103 more rows
## # ℹ 20 more variables: cf3marm <dbl>, cf3kids <dbl>, cf3md_case_lib <dbl>,
## #   cf3hhinc <dbl>, f3c3c <dbl>, m3c3c <dbl>, cf4cohm <dbl>, t4d7 <dbl>,
## #   p4l63 <dbl>, p4l59 <dbl>, p4d1c <dbl>, m4d1b <dbl>, f4b8d <dbl>,
## #   k5d1f <dbl>, k5d1g <dbl>, k5h1 <dbl>, k5d1e <dbl>, k5f1f <dbl>,
## #   y_missing_rate <dbl>, y_score <dbl>
```

```
# Father baseline education (father report, then mother report)
library(ggplot2)
# Father
# Filter the data to exclude rows where cfledu equals -3
filtered_data <- data[data$cfledu != -3, ]
# Create a bar plot
ggplot(filtered_data, aes(x = cfledu, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=0.9)  +
  scale_fill_gradient(low='#808000', high='#FF7F50') +
  labs(title = "Distribution of cfledu")
```

```
## Warning: The dot-dot notation (`..x..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(x)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Mother
# Filter the data to exclude rows where cfledu equals -3
filtered_data_m <- data[data$cmledu != -3, ]
# Create a bar plot
ggplot(filtered_data_m, aes(x = cmledu, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=0.9)  +
  scale_fill_gradient(low='#483D8B', high='#DB7093') +
  labs(title = "Distribution of cmledu")
```

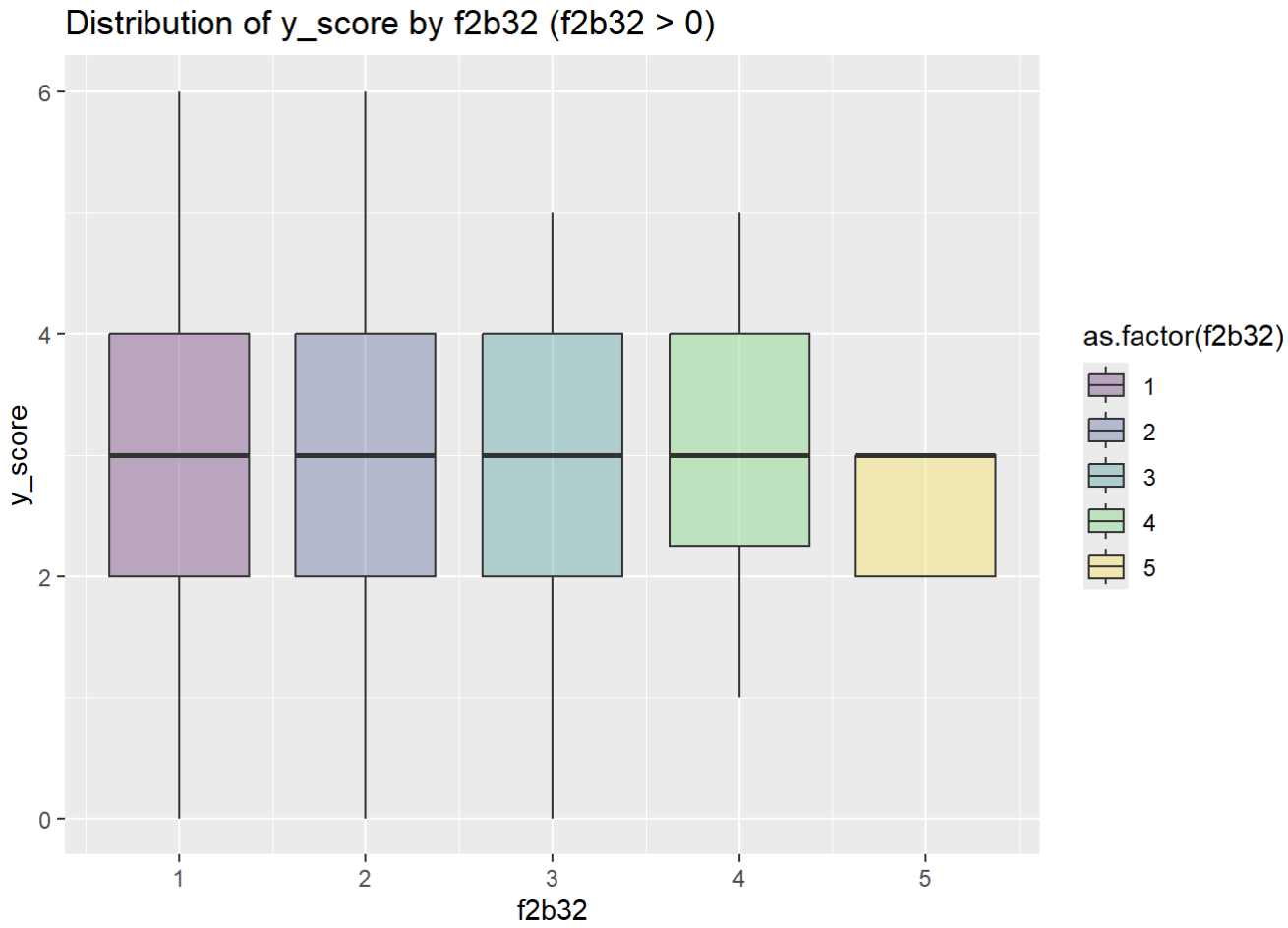



```
# f2b13     Does child walk or crawl yet?
```

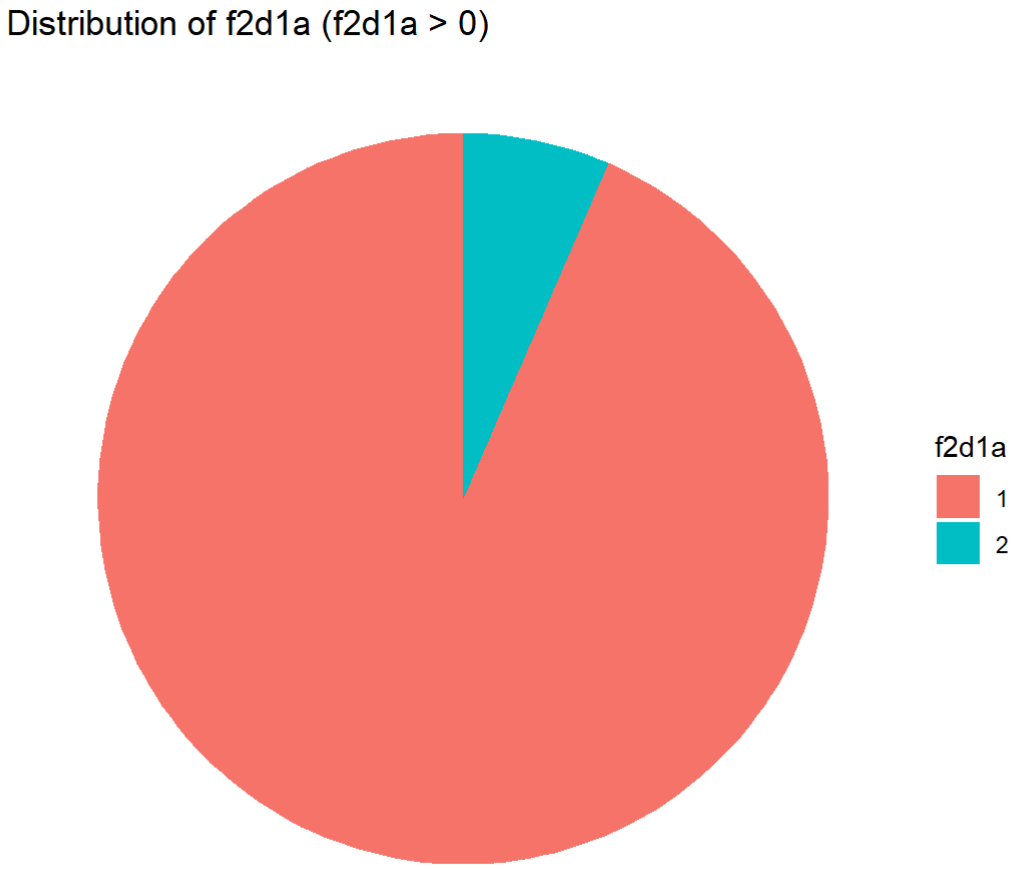
```
# f2b32       In general, how is your child's health?
library(viridis)
```

```
## 载入需要的程辑包：viridisLite
```

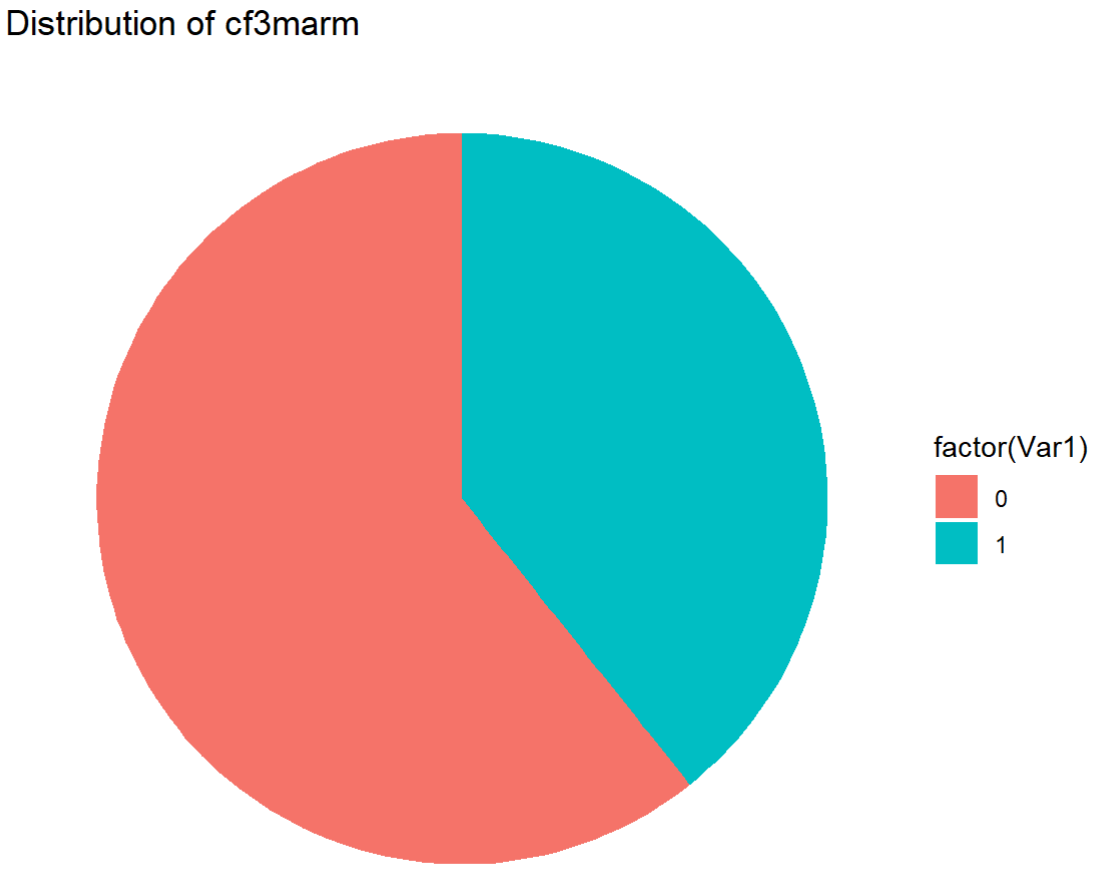
```
# Filter the data without negative
filtered_data_h <- data[data$f2b32 > 0, ]
# boxplot
ggplot(filtered_data_h, aes(y = y_score, x = f2b32, fill = as.factor(f2b32))) +
  geom_boxplot(alpha = 0.3)  +
  scale_fill_viridis_d() +
  labs(title = "Distribution of y_score by f2b32 (f2b32 > 0)")
```



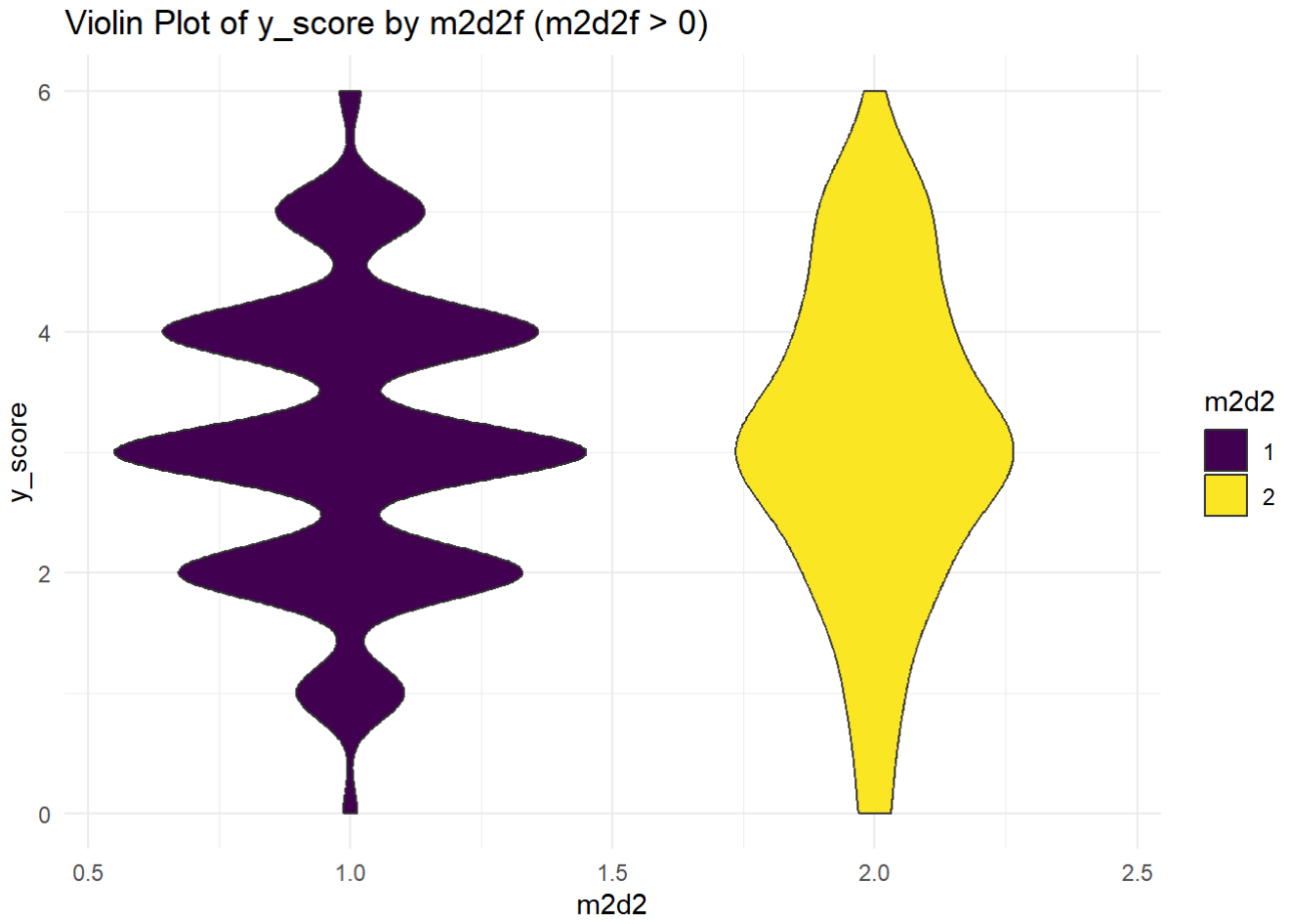

```
# yes or no
# f2d1a      Int Chk: Does mother have any contact with child?
# Filter the data without negative
filtered_data <- data[data$f2d1a >= 0, ]
# Count the frequency of each category in f2d1a
category_counts <- table(filtered_data$f2d1a)
# Convert the frequency table to a data frame
category_df <- as.data.frame(category_counts)
# Create a pie chart
ggplot(category_df, aes(x = "", y = Freq, fill = factor(Var1))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of f2d1a (f2d1a > 0)", fill = "f2d1a") +
  theme_void() +
  theme(legend.position = "right")
```



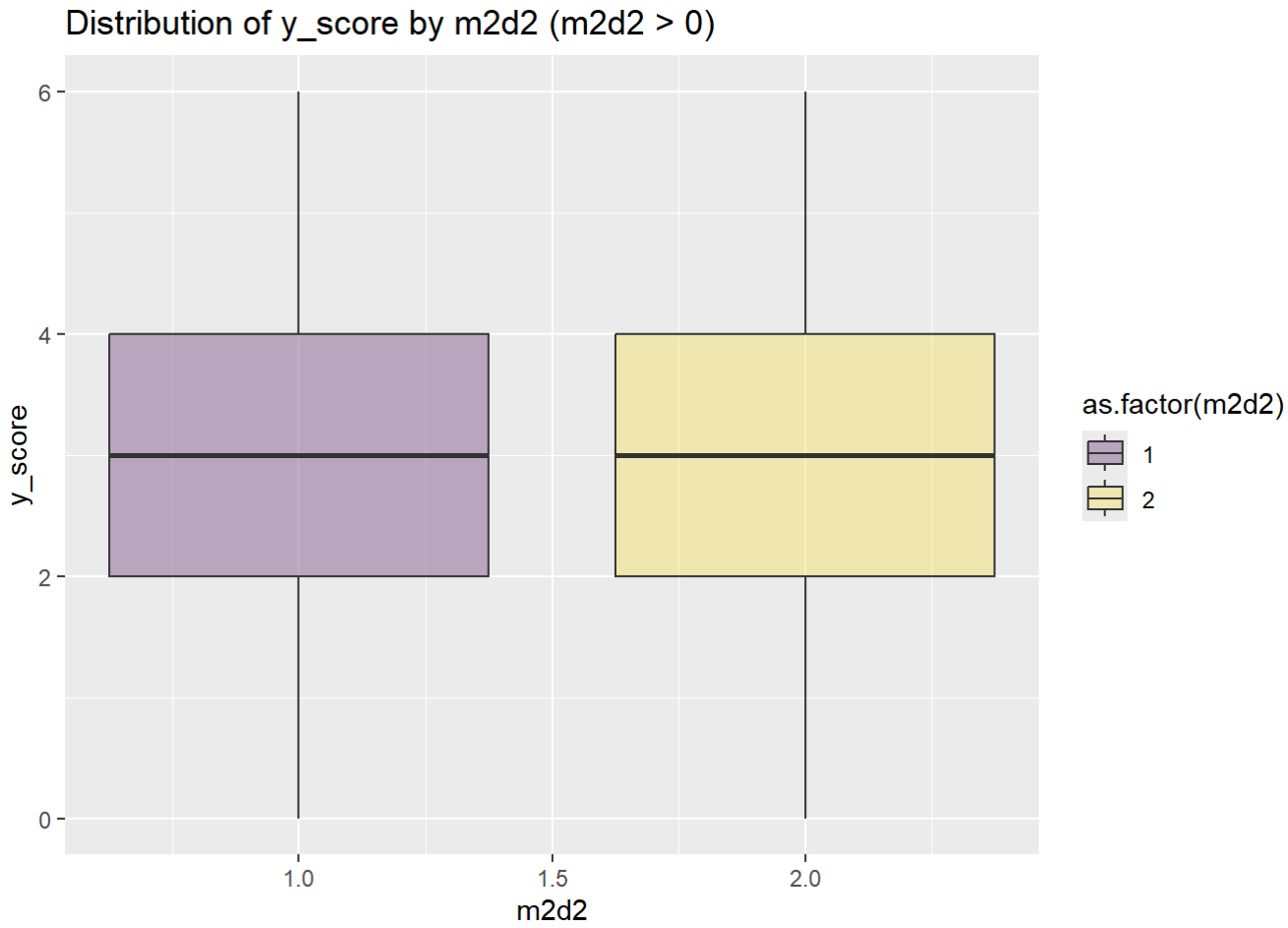

```
# cf3marm      Constructed - Is father married to child's mother at year three?  - wave3
# Filter the data without negative
filtered_data <- data[data$cf3marm >= 0, ]
# Count the frequency of each category in f2d1a
category_counts <- table(filtered_data$cf3marm)
# Convert the frequency table to a data frame
category_df <- as.data.frame(category_counts)
# Create a pie chart
ggplot(category_df, aes(x = "", y = Freq, fill = factor(Var1))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of cf3marm") +
  theme_void() +
  theme(legend.position = "right")
```



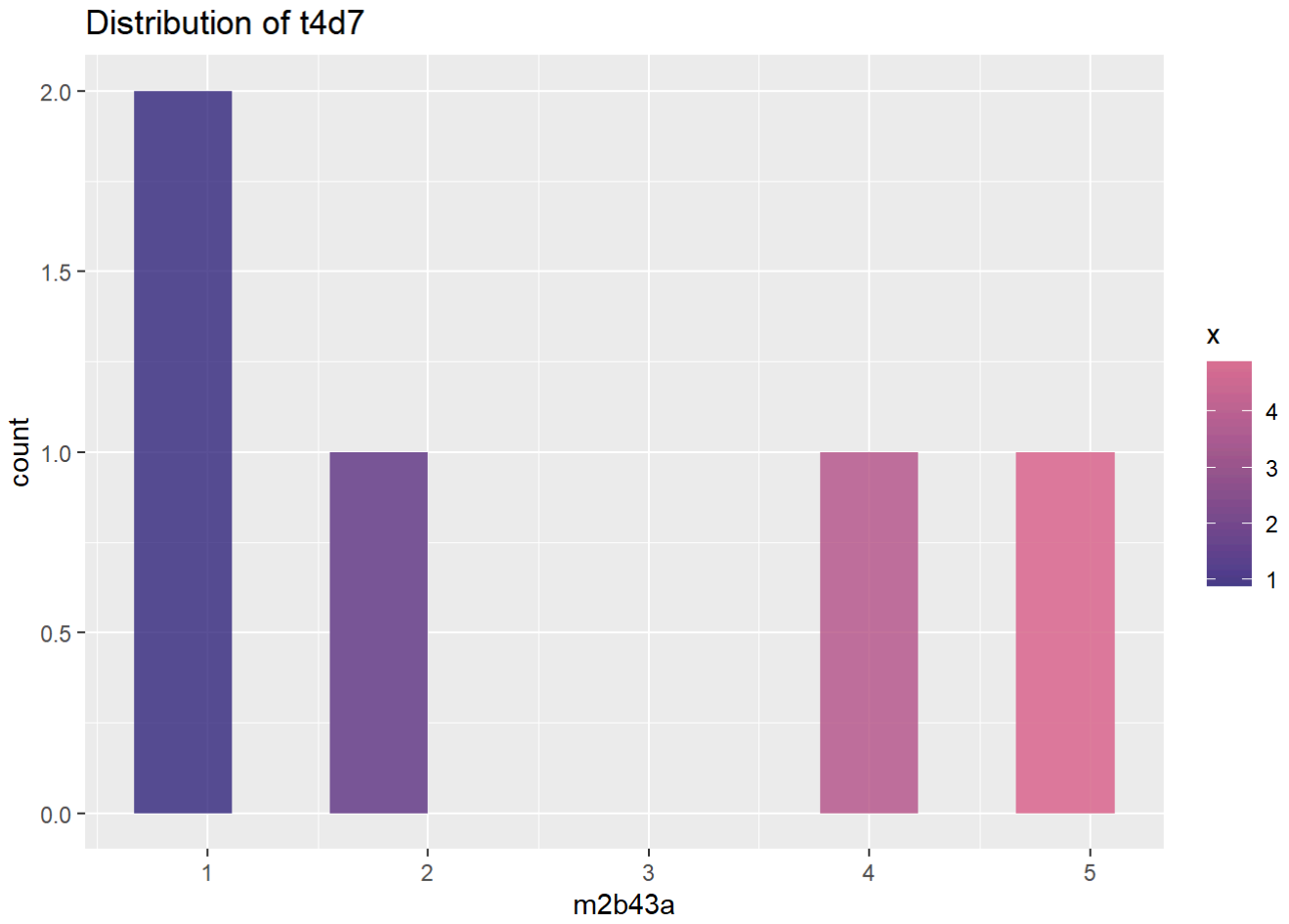

```
# m2d2f      How often-Can count on father to watch child for a few hours?
# maybe wrong
# Filter the data without negative
filtered_data <- data[data$m2d2 > 0, ]
# violin plot
ggplot(filtered_data, aes(x = m2d2, y = y_score, fill = as.factor(m2d2))) +
  geom_violin() +
  scale_fill_viridis_d() +
  labs(title = "Violin Plot of y_score by m2d2f (m2d2f > 0)", x = "m2d2", y = "y_score", fill =
"m2d2") +
  theme_minimal()
```



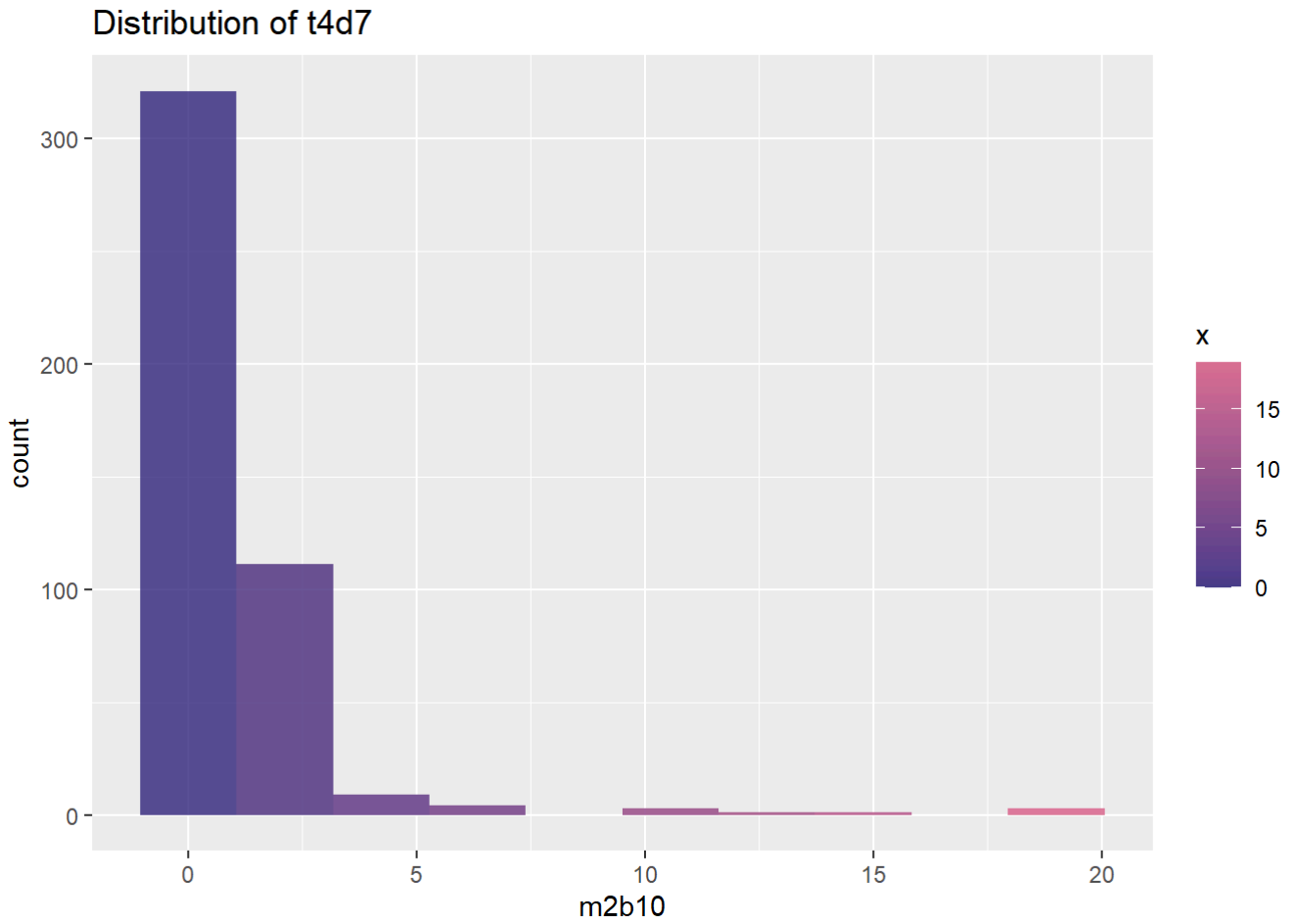

```
# boxplot
ggplot(filtered_data, aes(y = y_score, x = m2d2, fill = as.factor(m2d2))) +
  geom_boxplot(alpha = 0.3)  +
  scale_fill_viridis_d() +
  labs(title = "Distribution of y_score by m2d2 (m2d2 > 0)")
```



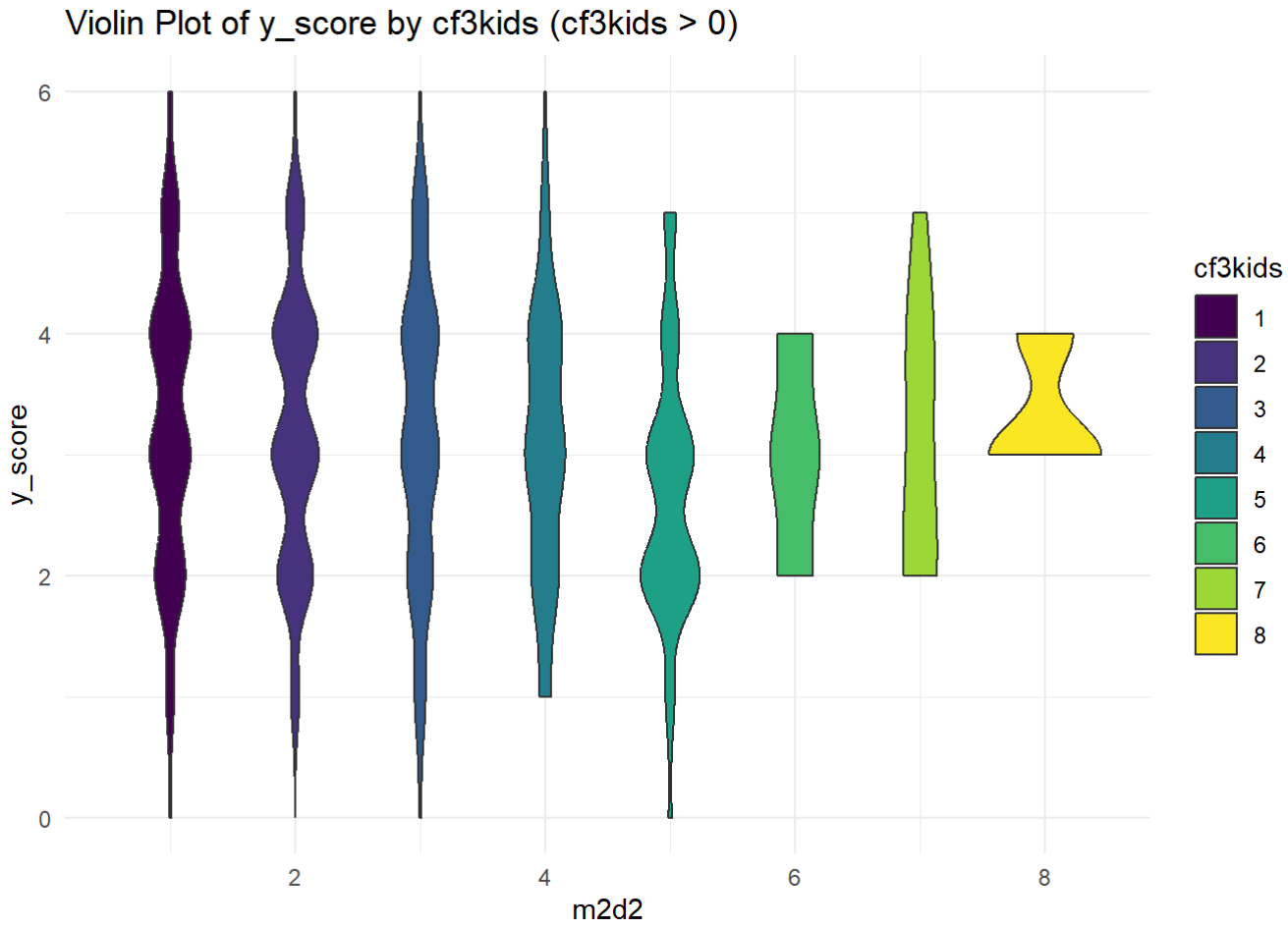

```
# wrong data
# m2b43a    On a scale of 1-(least like) to 5-(most like) - Child tends to be shy
# Filter the data to exclude rows where cfledu equals -3
filtered_data_m <- data[data$m2b43a >= 0, ]
# Create a bar plot
ggplot(filtered_data_m, aes(x = m2b43a, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=0.9)  +
  scale_fill_gradient(low='#483D8B', high='#DB7093') +
  labs(title = "Distribution of t4d7")
```



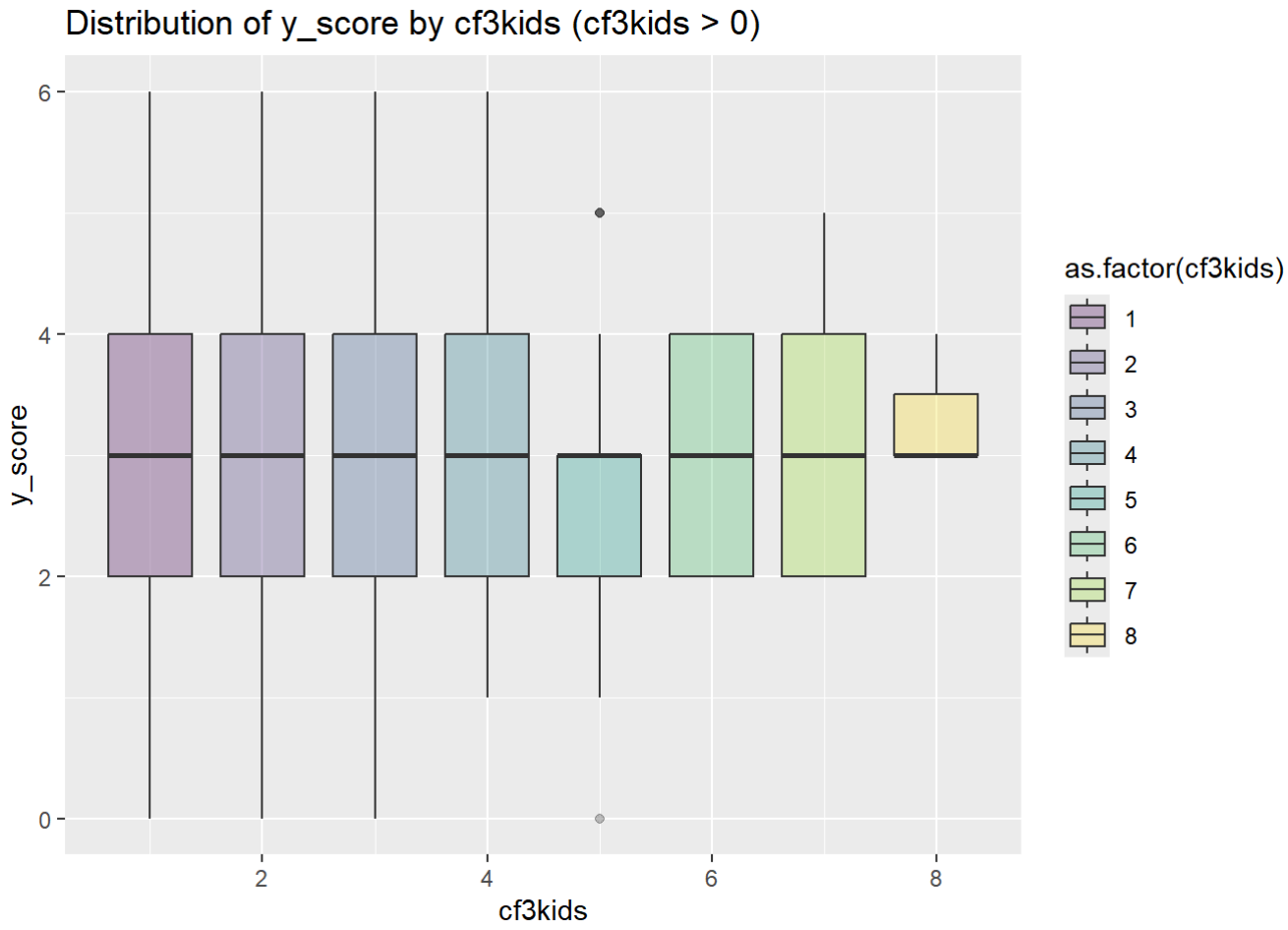

```
# wrong label
# m2b10b   How long did child stay in the hospital during longest stay (days)?
# Filter the data to exclude rows where cfledu equals -3
filtered_data_n <- data[data$m2b10 > 0, ]
# Create a bar plot
ggplot(filtered_data_n, aes(x = m2b10, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=0.9)  +
  scale_fill_gradient(low='#483D8B', high='#DB7093') +
  labs(title = "Distribution of t4d7")
```



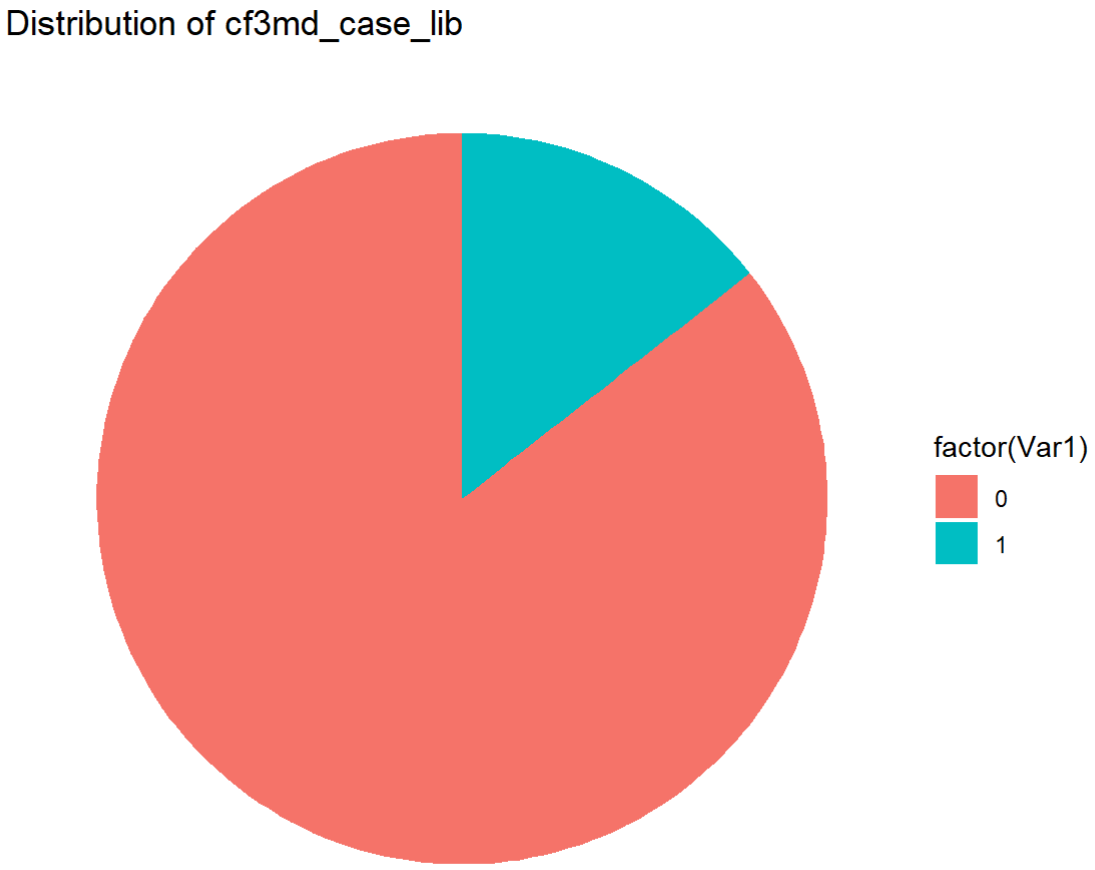

```
# cf3kids   Constructed - Number of children under 18 in household
# Filter the data without negative
filtered_data <- data[data$cf3kids > 0, ]
# violin plot
ggplot(filtered_data, aes(x = cf3kids, y = y_score, fill = as.factor(cf3kids))) +
  geom_violin() +
  scale_fill_viridis_d() +
  labs(title = "Violin Plot of y_score by cf3kids (cf3kids > 0)", x = "m2d2", y = "y_score", fi
ll = "cf3kids") +
  theme_minimal()
```



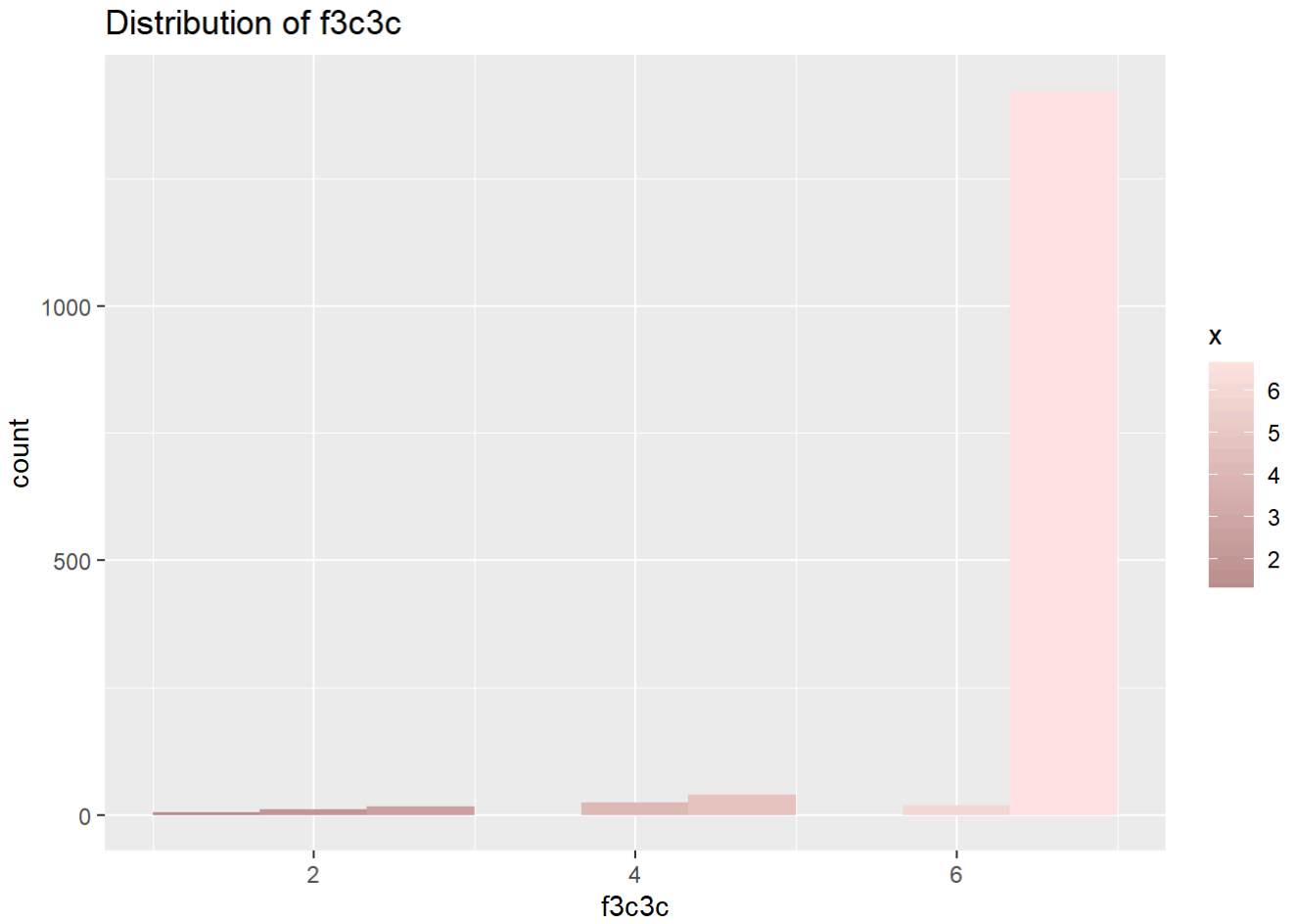

```
# boxplot
ggplot(filtered_data, aes(y = y_score, x = cf3kids, fill = as.factor(cf3kids))) +
  geom_boxplot(alpha = 0.3)  +
  scale_fill_viridis_d() +
  labs(title = "Distribution of y_score by cf3kids (cf3kids > 0)")
```



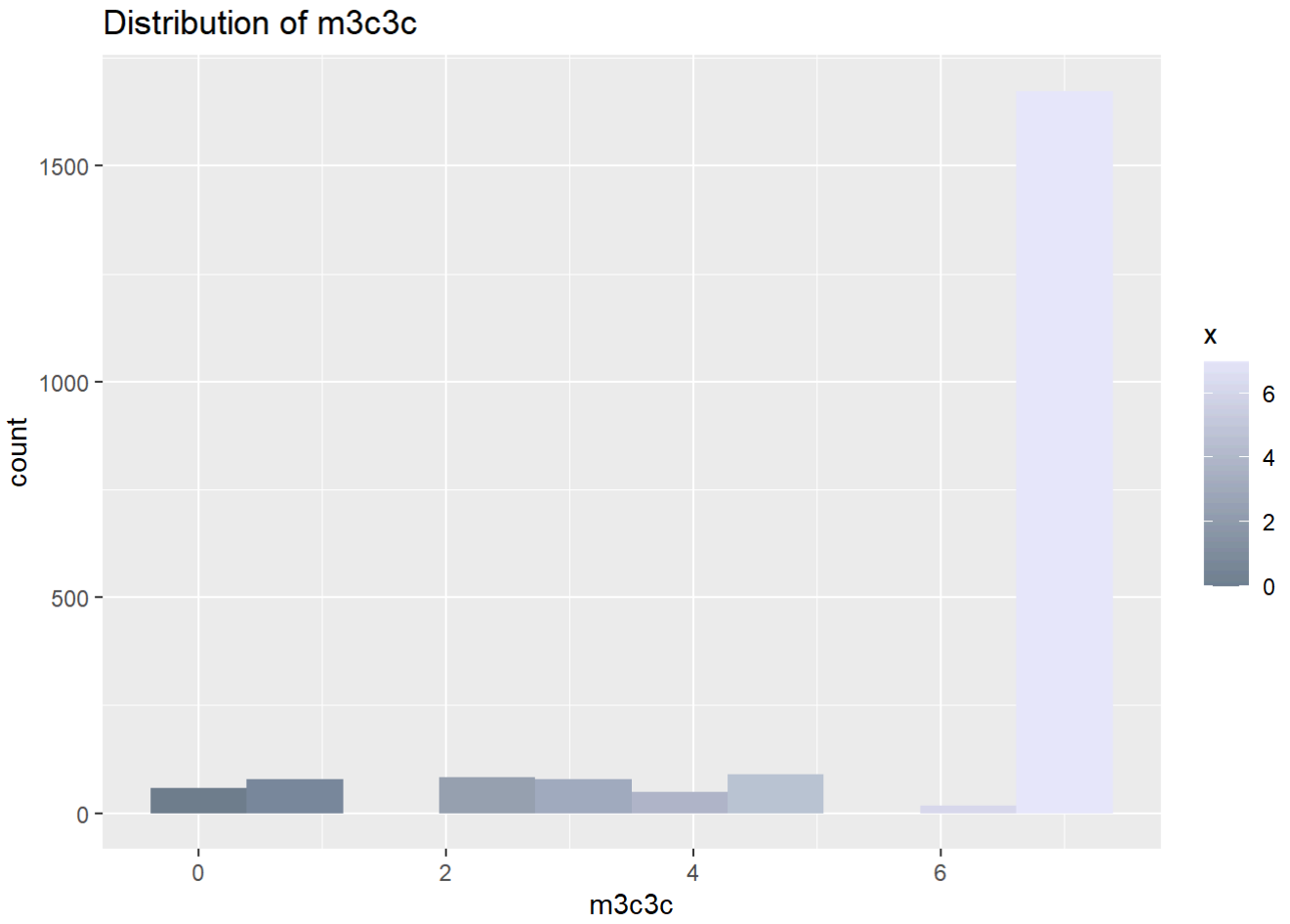

```
# cf3md_case_lib        Constructed - Father meets depression criteria (liberal) at three-year (CIDI)
# Filter the data without negative
filtered_data <- data[data$cf3md_case_lib >= 0, ]
# Count the frequency of each category in f2d1a
category_counts <- table(filtered_data$cf3md_case_lib)
# Convert the frequency table to a data frame
category_df <- as.data.frame(category_counts)
# Create a pie chart
ggplot(category_df, aes(x = "", y = Freq, fill = factor(Var1))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of cf3md_case_lib") +
  theme_void() +
  theme(legend.position = "right")
```



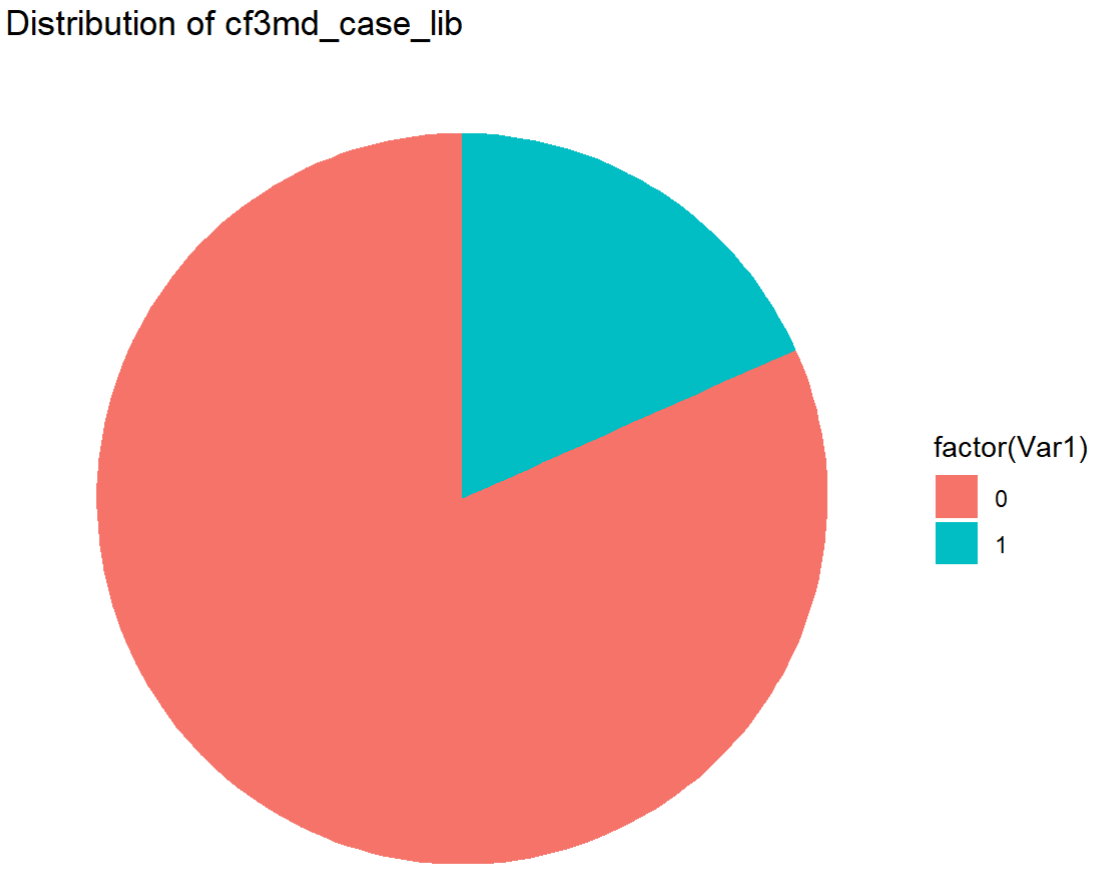

```
# f3c3c      Days/week: tell child she loves him/her?
# Filter the data to exclude rows where cfledu equals -3
filtered_data_m <- data[data$f3c3c > 0, ]
# Create a bar plot
ggplot(filtered_data_m, aes(x = f3c3c, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=1)  +
  scale_fill_gradient(low='#BC8F8F', high='#FFE4E1') +
  labs(title = "Distribution of f3c3c")
```



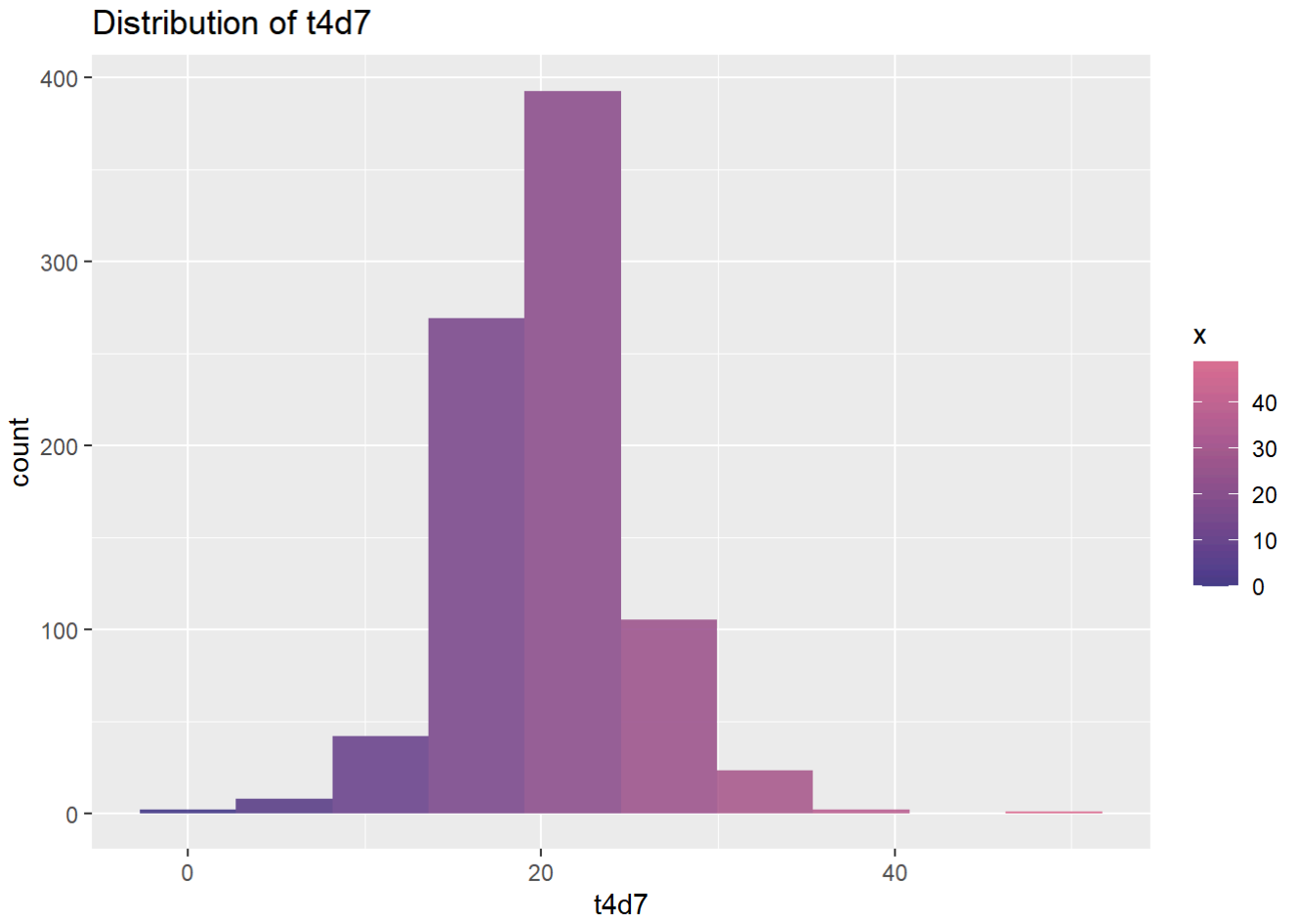

```
# m3c3c         Days/week: tell child he loves him/her?
# Filter the data to exclude rows where cfledu equals -3
filtered_data_m <- data[data$m3c3c >= 0, ]
# Create a bar plot
ggplot(filtered_data_m, aes(x = m3c3c, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=1)  +
  scale_fill_gradient(low='#708090', high='#E6E6FA') +
  labs(title = "Distribution of m3c3c")
```




```
# cf4cohm        Constructed - Father living with child's mother at five-year
# Filter the data without negative
filtered_data <- data[data$cf4cohm >= 0, ]
# Count the frequency of each category in f2d1a
category_counts <- table(filtered_data$cf4cohm)
# Convert the frequency table to a data frame
category_df <- as.data.frame(category_counts)
# Create a pie chart
ggplot(category_df, aes(x = "", y = Freq, fill = factor(Var1))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of cf3md_case_lib") +
  theme_void() +
  theme(legend.position = "right")
```




```
# t4d7          -->d7. number of kids present with child
# Filter the data to exclude rows where cfledu equals -3
filtered_data_m <- data[data$t4d7 >= 0, ]
# Create a bar plot
ggplot(filtered_data_m, aes(x = t4d7, fill = ..x..)) +
  geom_histogram(bins = 10, alpha=0.9)  +
  scale_fill_gradient(low='#483D8B', high='#DB7093') +
  labs(title = "Distribution of t4d7")
```




```
# Add a line chart
#geom_line(stat = "count", aes(y = ..count.. * 50, group = 1), color = "red") +
# Adjust the y-axis scale for the line chart
#scale_y_continuous(sec.axis = sec_axis(~./50, name = "Line chart")) +
# Specify the legend for the line chart
# guides(fill = guide_legend(title = "Bar plot"), color = guide_legend(title = "Line chart"))
```