Homework 1

Chi Zhang UNI:cz2481

Jan 30, 2018

For all questions involving histograms, choose a sensible binwidth and breakpoints, unless otherwise indicated.

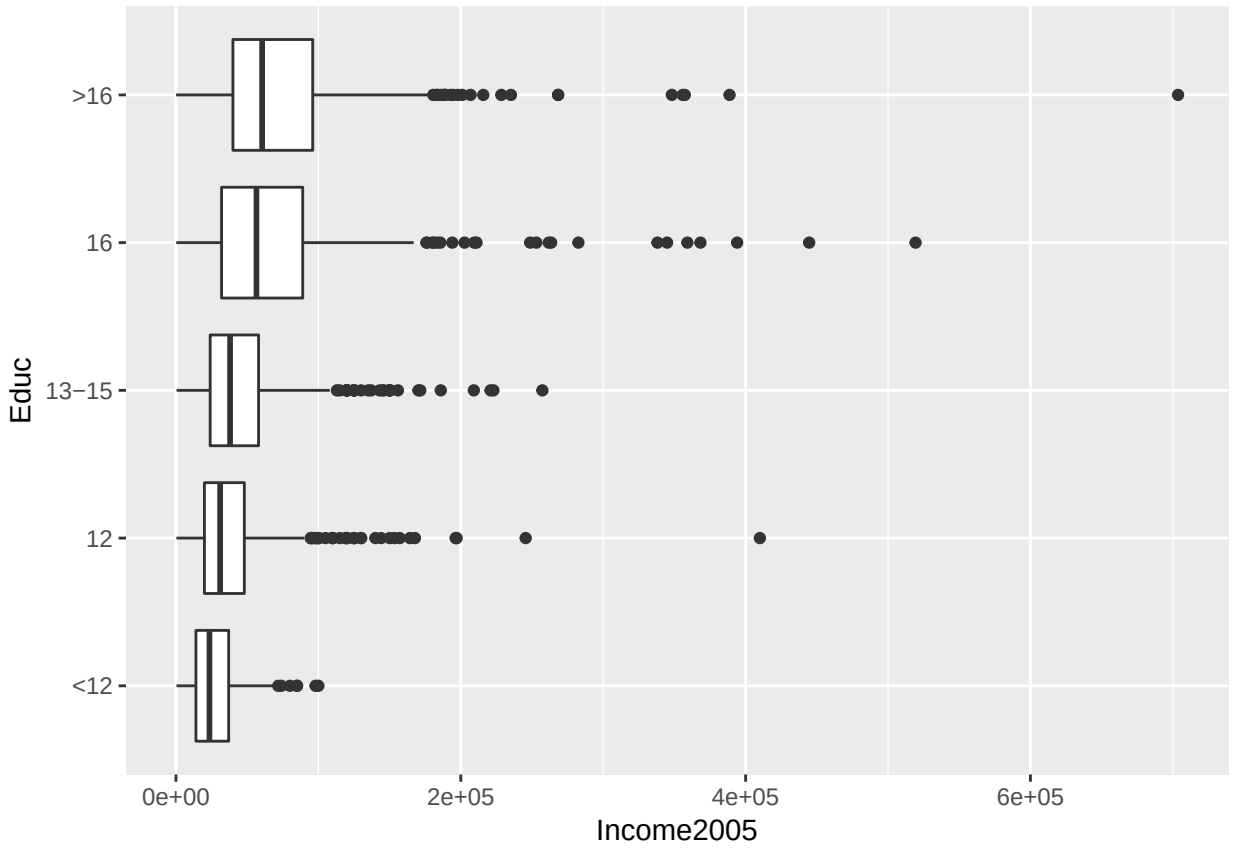
1. Income

a) Describe in detail the features you observe in the boxplots below, plotted with data from the *ex0525* dataset, **Sleuth3** page. (see page 29 in *Graphical Data Analysis in R* for a list of features to concentrate on, and the numbered list on the bottom of page 43 for an example of how to describe features of a graph in words.) [5 points]

```
install.packages("Sleuth3",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4w/fp4pnk193ls5y5byhtt_8smw0000gn/T//RtmpemXuV2/downloaded_packages
```

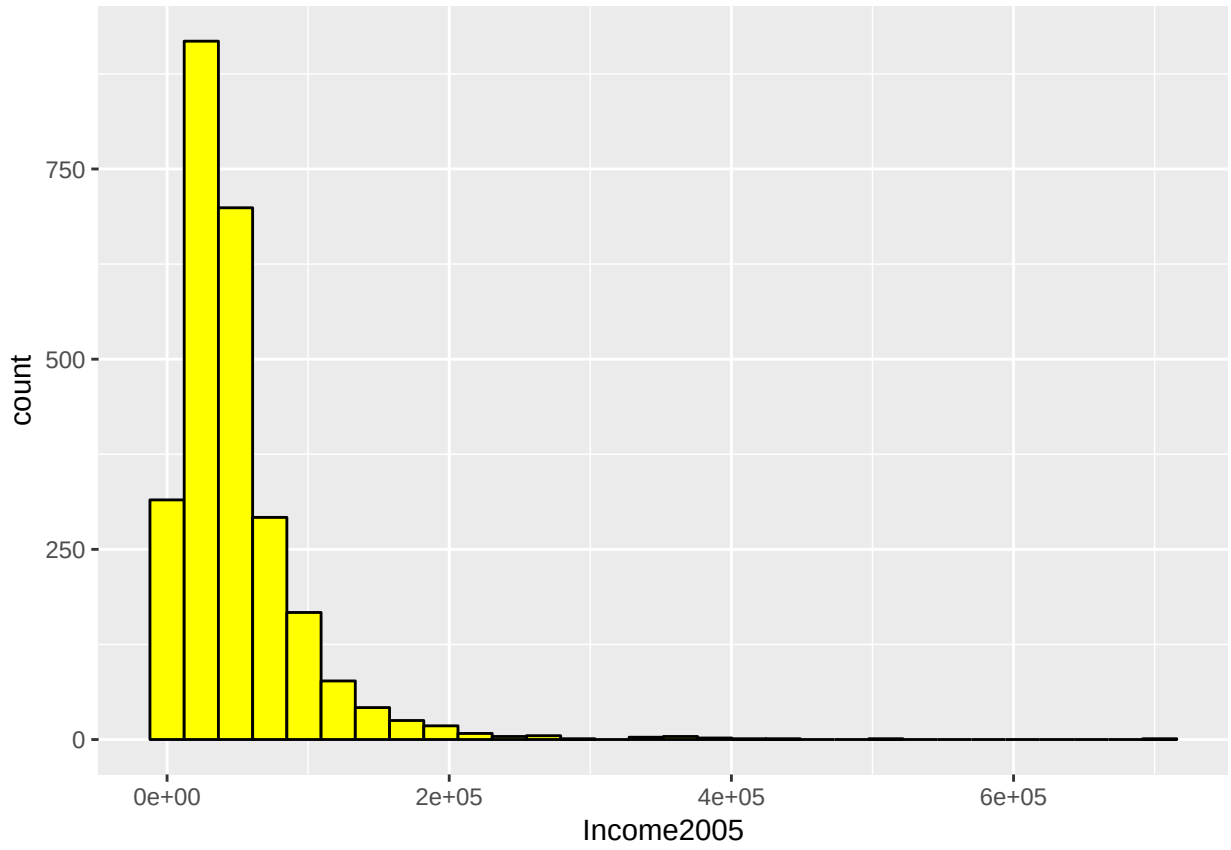
```
library(Sleuth3)
library(tidyverse)

# convert Educ from an integer to a factor, and make "<12" the first factor level
mydata <- ex0525 %>%
  dplyr::mutate(Educ = forcats::fct_relevel(Educ, "<12"))

ggplot(mydata, aes(Educ, Income2005)) +
  geom_boxplot() +
  coord_flip()   # for horizontal boxplots
```



1. There is serveral obvious outliers in group '12' and '>16'.
2. There is no clear outlier in group '<12'.
3. There is a higher median if the group has a higher education level.
4. There is a biger IQR if the group has a higher education level.

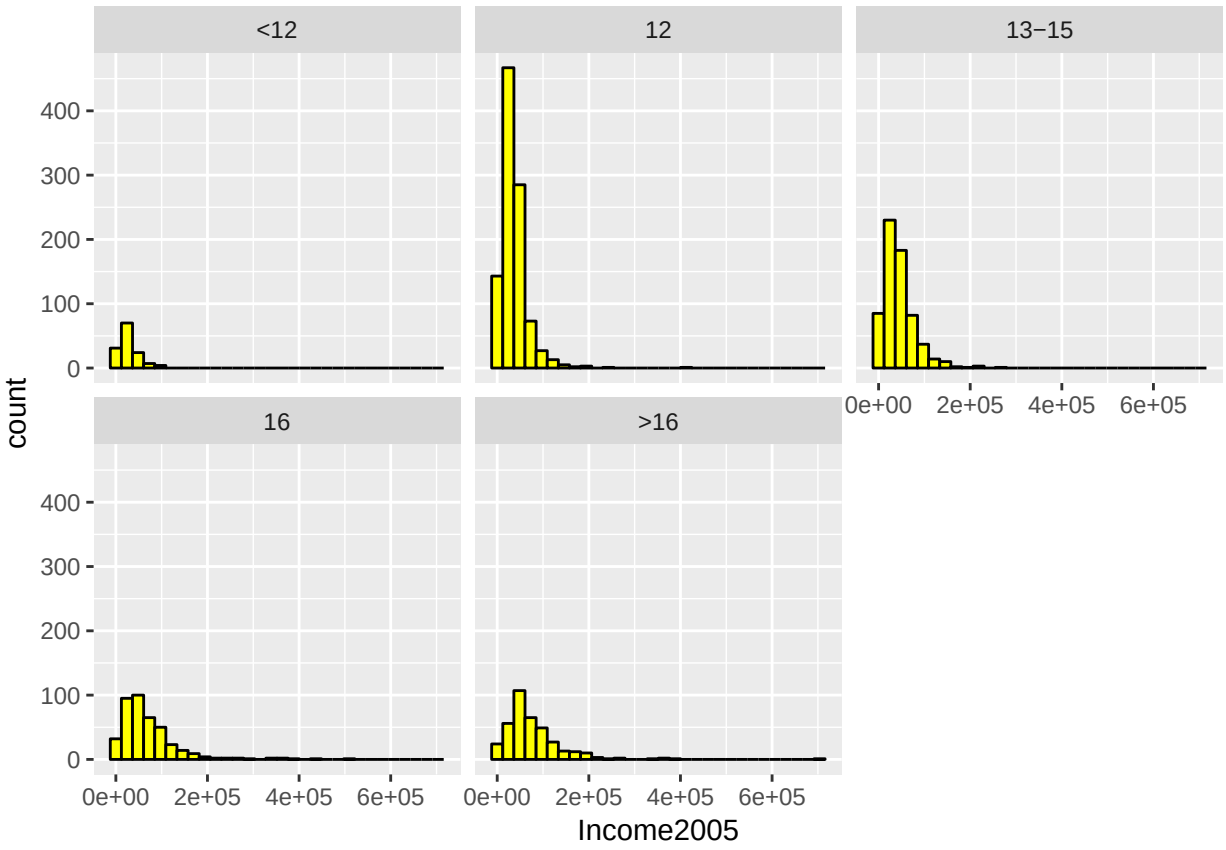
b) Plot a histogram of the `Income2005` variable in the dataset referenced in part a). [3 points]

```
ggplot(mydata, aes(Income2005)) + geom_histogram(color = "black", fill = "yellow")
```



c) Use `+facet_wrap(~Educ)` to facet the histogram on education level. [3 points]

```
ggplot(mydata, aes(Income2005)) + geom_histogram(color = "black", fill = "yellow") + facet_wrap(~Educ)
```



d) What do you learn from the histograms that wasn't apparent in the boxplots from question 1? [3 points]

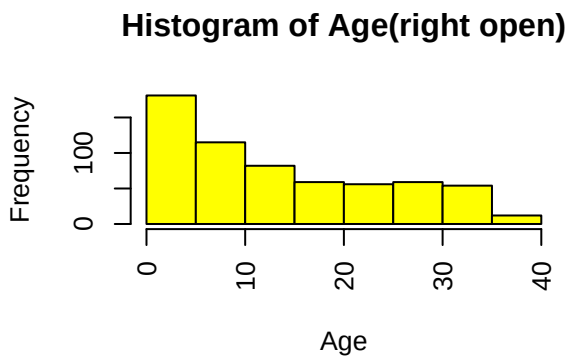
The size of the different age band dataset. '12' and '13-15' have relatively larger size than other three. And the '<12' has smallest dataset size.

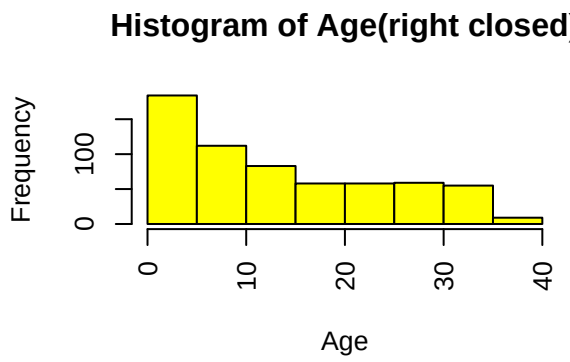
2. Respiratory Rates

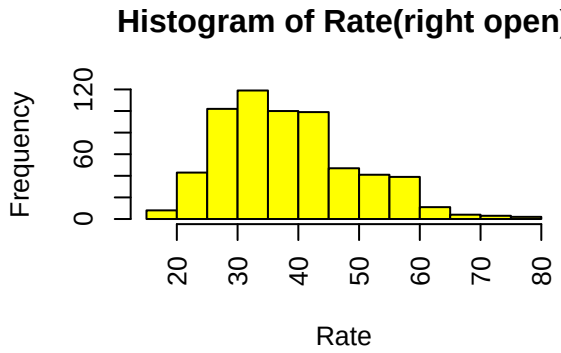
a) Plot right closed and right open histograms for each of the two variables in the *ex0824* dataset in the **Sleuth3** package using default binwidths and breaks. (4 histograms in total). [4 points]

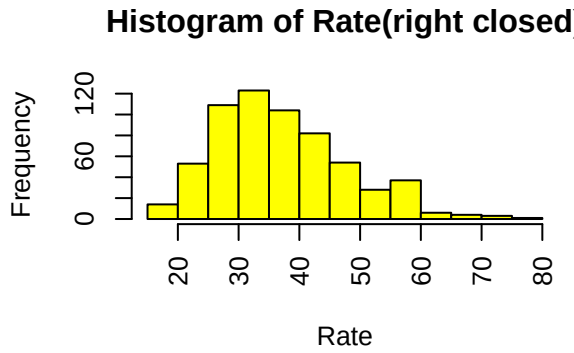
1.

```
mydata <- ex0824
par(mfrow = c(2, 2), las = 3)
Histogram = hist(mydata$Age, col = "yellow", right = FALSE, xlab = 'Age', main = 'Histogram of Age(right
hist(mydata$Age, col = "yellow", left = FALSE, xlab = 'Age', main = 'Histogram of Age(right closed)' )
hist(mydata$Rate, col = "yellow", right = FALSE, xlab = 'Rate', main = 'Histogram of Rate(right open)')
hist(mydata$Rate, col = "yellow", left = FALSE, xlab = 'Rate', main = 'Histogram of Rate(right closed)'
```







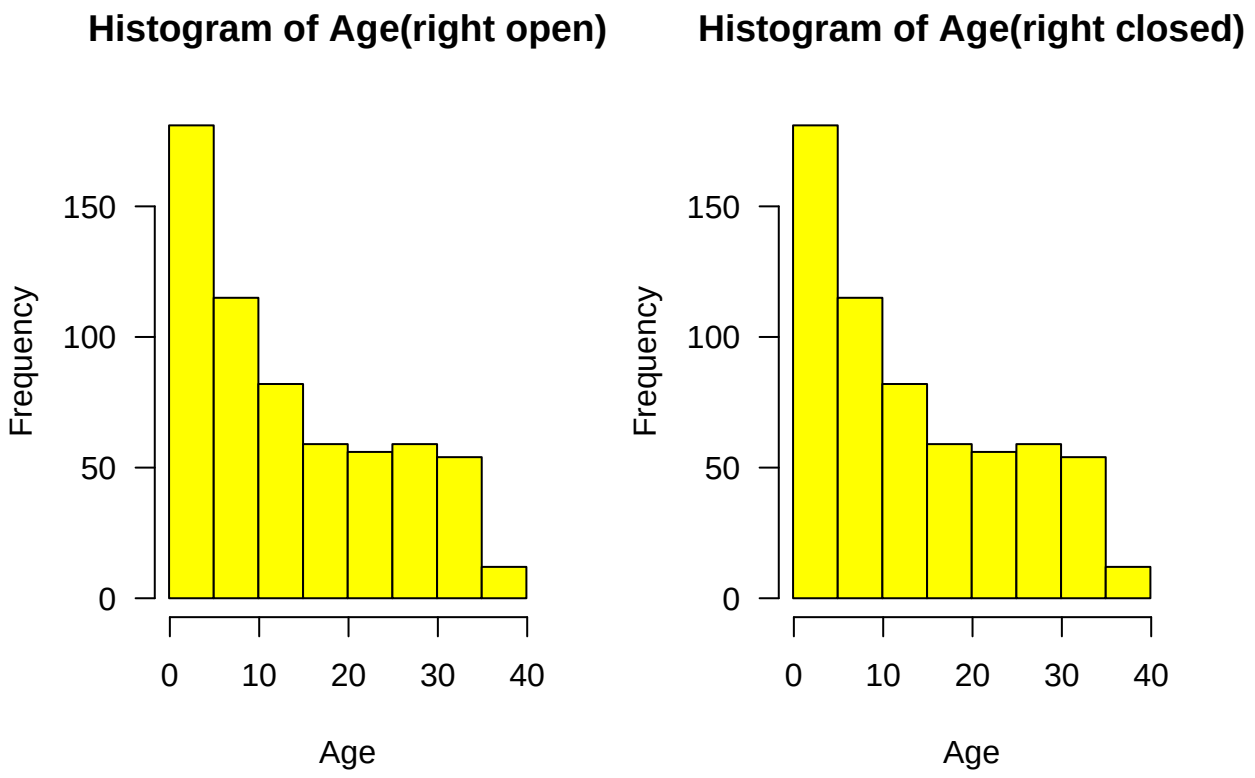


b) For which variable, `Age` or `Rate`, do the two versions differ more? Why? [3 points]

"Age" differs more since "Age" data are decimals, the probability that "Age" data point falls on the boundary would much lower than "Rate"(intgers data).

c) Redraw the `Age` histograms with different parameters so that the right closed and right open versions are identical. [3 points]

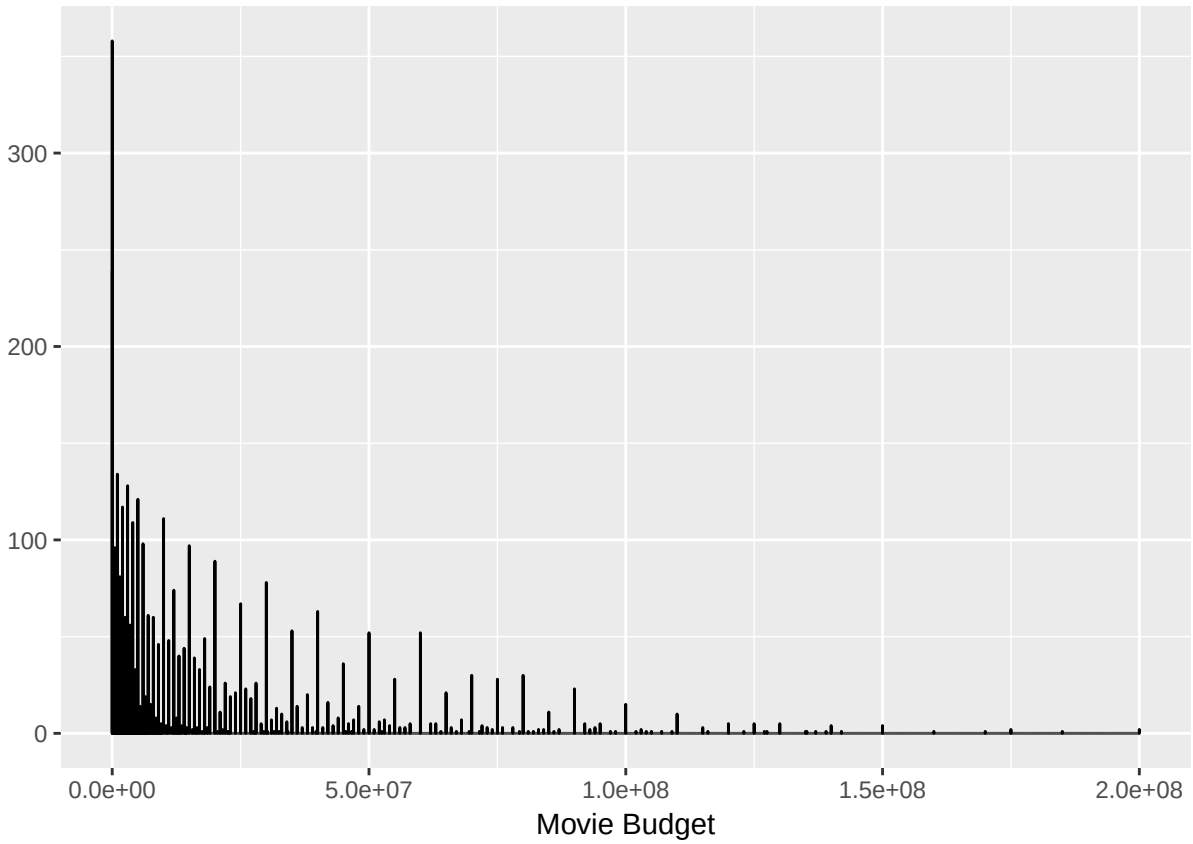
```
par(mfrow = c(1, 2), las = 1)
hist(mydata$Age, col = "yellow", breaks = Histogram$breaks - 0.09, right = FALSE, xlab = 'Age', main =
hist(mydata$Age, col = "yellow", breaks = Histogram$breaks - 0.09, left = FALSE, xlab = 'Age', main = '
```



3. Movie budgets

Are there rounding patterns in the `budget` variable of the *movies* in the **ggplot2movies** package? If so, what are the patterns? (Note: according to the textbook this dataset is in the **ggplot2** package, but it has since been moved to a separate package.) Support your conclusions with graphical evidence. You are encouraged to break the variable down into different budget ranges and consider them separately. [8 points]

```
library(ggplot2movies)
ggplot(movies, aes(budget)) + geom_histogram(color='black',bins =10000) +
xlab("Movie Budget") + ylab("")
```



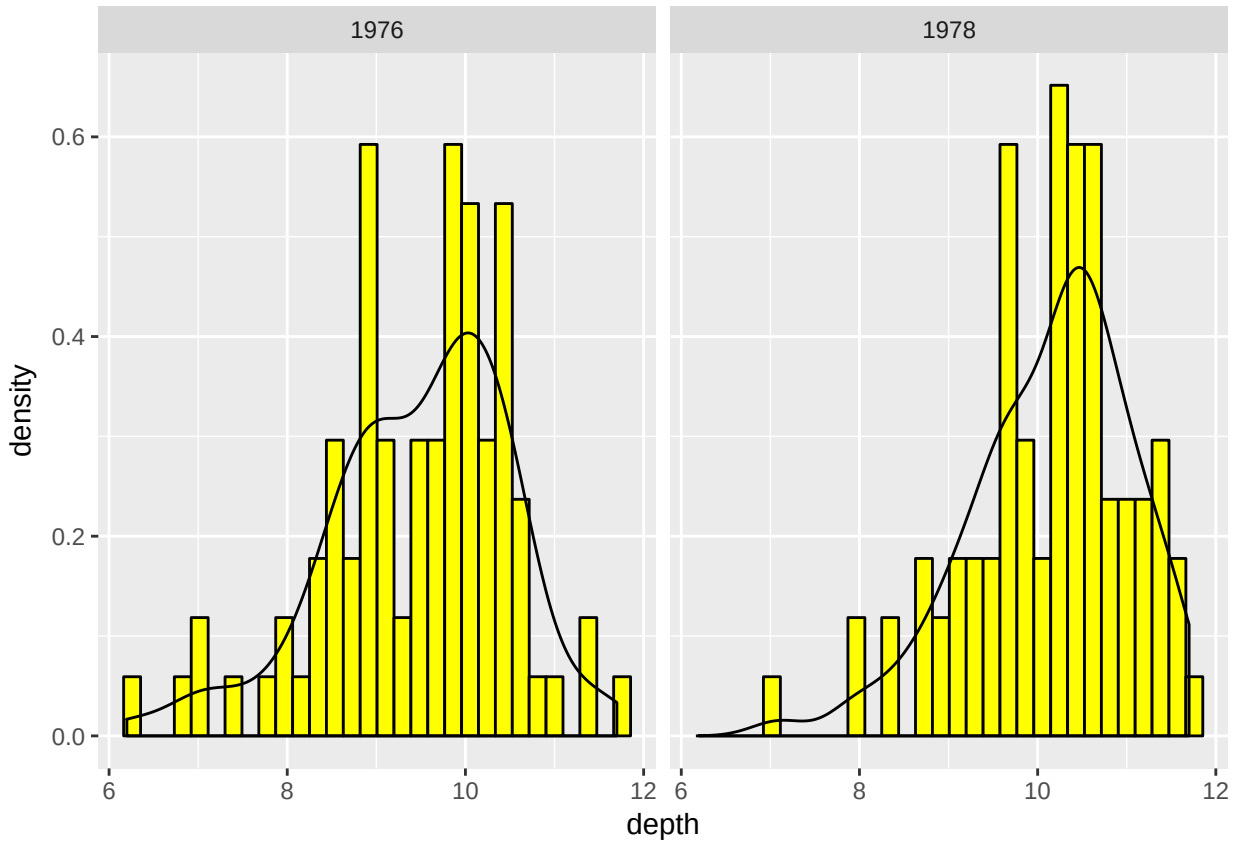
There are about 10 peaks from 0 to 5e+7 and 10 peaks within every 5e+7. Thus, the rounding pattern could be:

1. For movies whose budget is less than 5e+7, the budget values often rounded to nearest 5e+6.
2. For movies whose budget is larger than 5e+7, the budget values often rounded to nearest 5e+7.

4. Finches

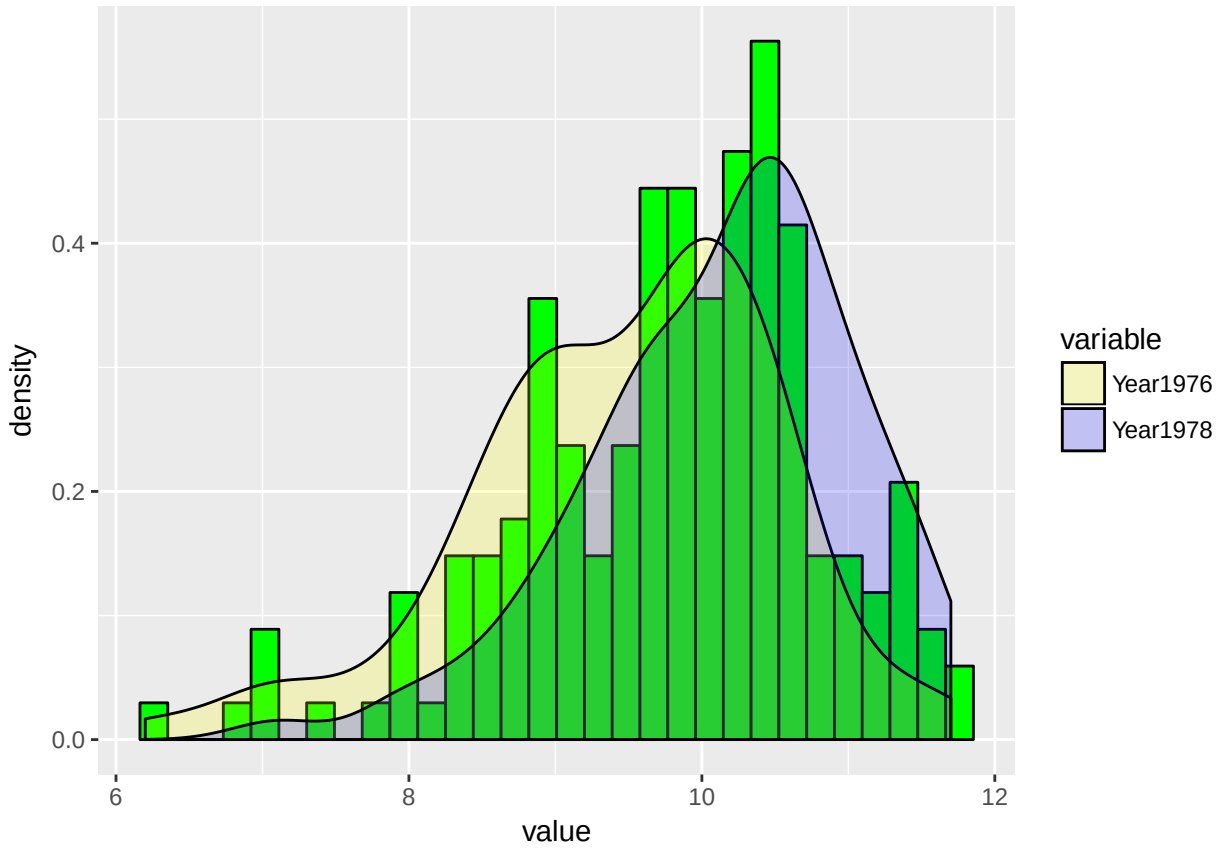
a) Plot separate density histograms of the beak depth of the finches in *case0201* from the **Sleuth3** package, with density curves overlaid as on page 34 of the textbook. (However, do this by facetting on `Year` rather than using `grid.arrange`). [3 points]

```
library(Sleuth3)
library(tidyverse)

# convert Educ from an integer to a factor, and make "<12" the first factor level
mydata <- case0201
depth <- mydata$Depth

ggplot(mydata, aes(depth)) + geom_histogram(aes(y=..density..),color='black',fill='yellow') + geom_dens
```



b) Plot both density curves on the same graph to facilitate comparison. Make 1976 yellow and 1978 blue. Use alpha blending so the fills are transparent. [3 points]

```
library(ggplot2)
library(reshape2)
depth <- mydata$Depth

data <- melt(data.frame(Year1976 = depth[mydata$Year == '1976'], Year1978 = depth[mydata$Year == '1978'])
ggplot(data,aes(value, fill=variable)) + geom_histogram(aes(y = ..density..), col = 'black',fill = 'gre
```



c) Based on your graphs in parts a) and b), describe how the distributions differ by year. [3 points]

The mean value trends to become larger in 1978 than 1976. And have a larger maximum value.

d) What is the cause of the difference according to the information in the help file? [3 points]

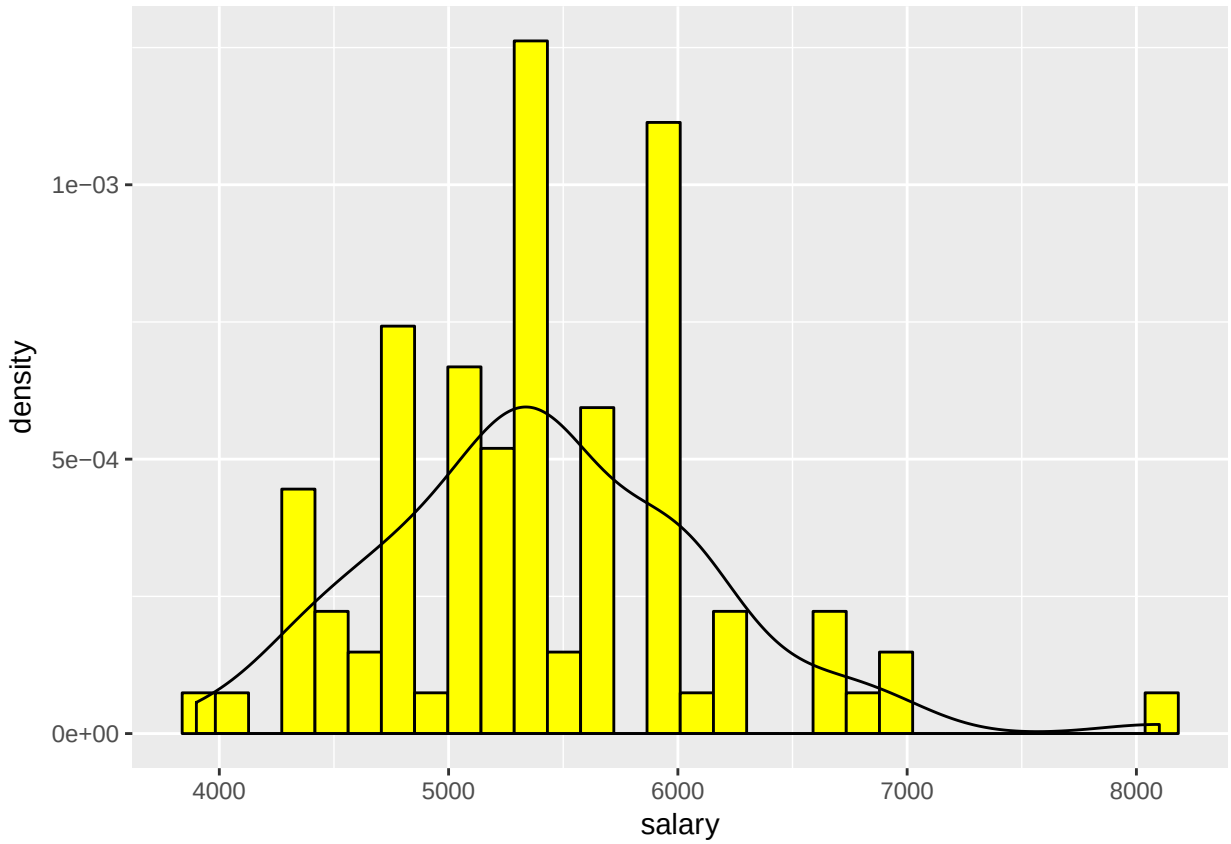
The finches have to open tough seeds for surviving, which requires a larger beak depths. And only those who has such a beak depth could survive to 1978.

5. Salary

Is the `Salary` variable in the *case0102* of **Sleuth3** normally distributed? Use two different graphical methods to provide evidence. [6 points]

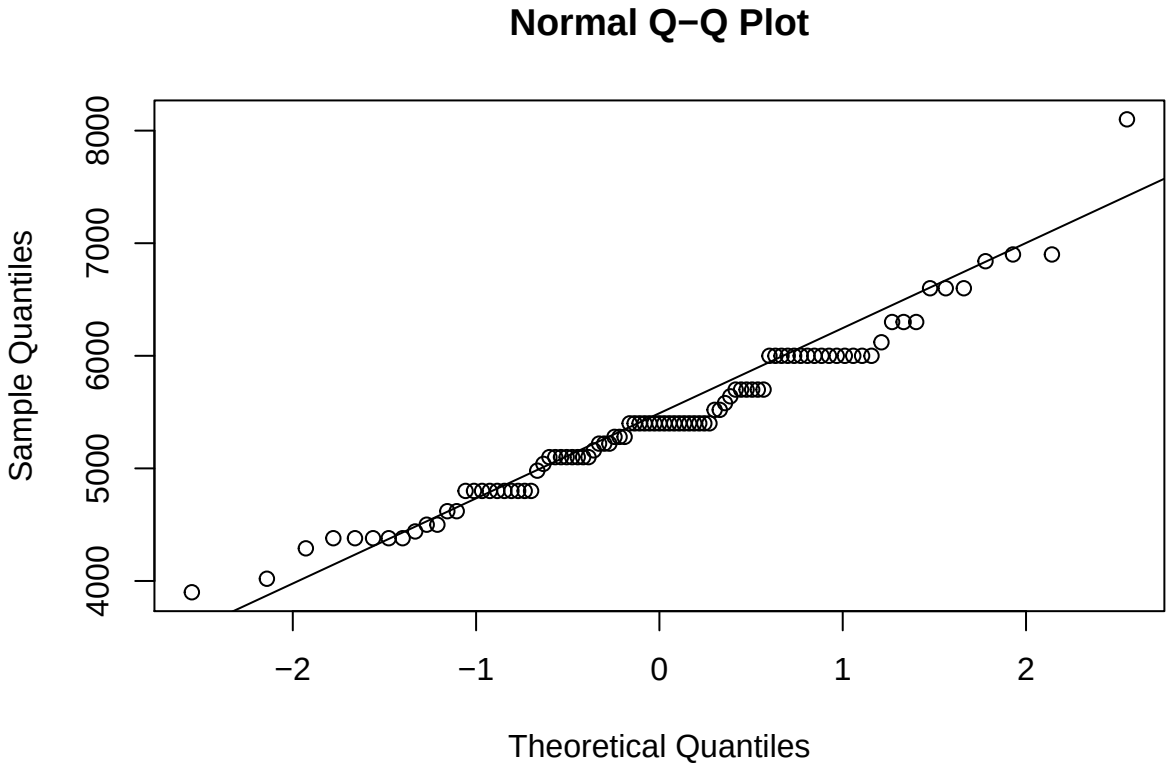
1. Draw the histogram of Salary. It's highly likely that the underlying distribution is not a normal distribution, since the density plot is not asymmetric.

```
library(Sleuth3)
library(tidyverse)

# convert Educ from an integer to a factor, and make "<12" the first factor level
mydata <- case0102
salary <- mydata$Salary

ggplot(mydata, aes(salary)) + geom_histogram(aes(y=..density..),color='black',fill='yellow') + geom_dens
```



2. Since the most of data don't tightly surround the diagonal of the first quadrant of Normal Q-Q Plot. The underlying distribution may not follow the normal distribution.

```
library(Sleuth3)
library(tidyverse)

# convert Educ from an integer to a factor, and make "<12" the first factor level
mydata <- case0102
qqnorm(mydata$Salary);qqline(mydata$Salary)
```


Normal Q−Q Plot
Sample Quantiles
4000
5000
6000
7000
8000
−2
−1
0
1
2
Theoretical Quantiles