R plots

Ikya Jupudy

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

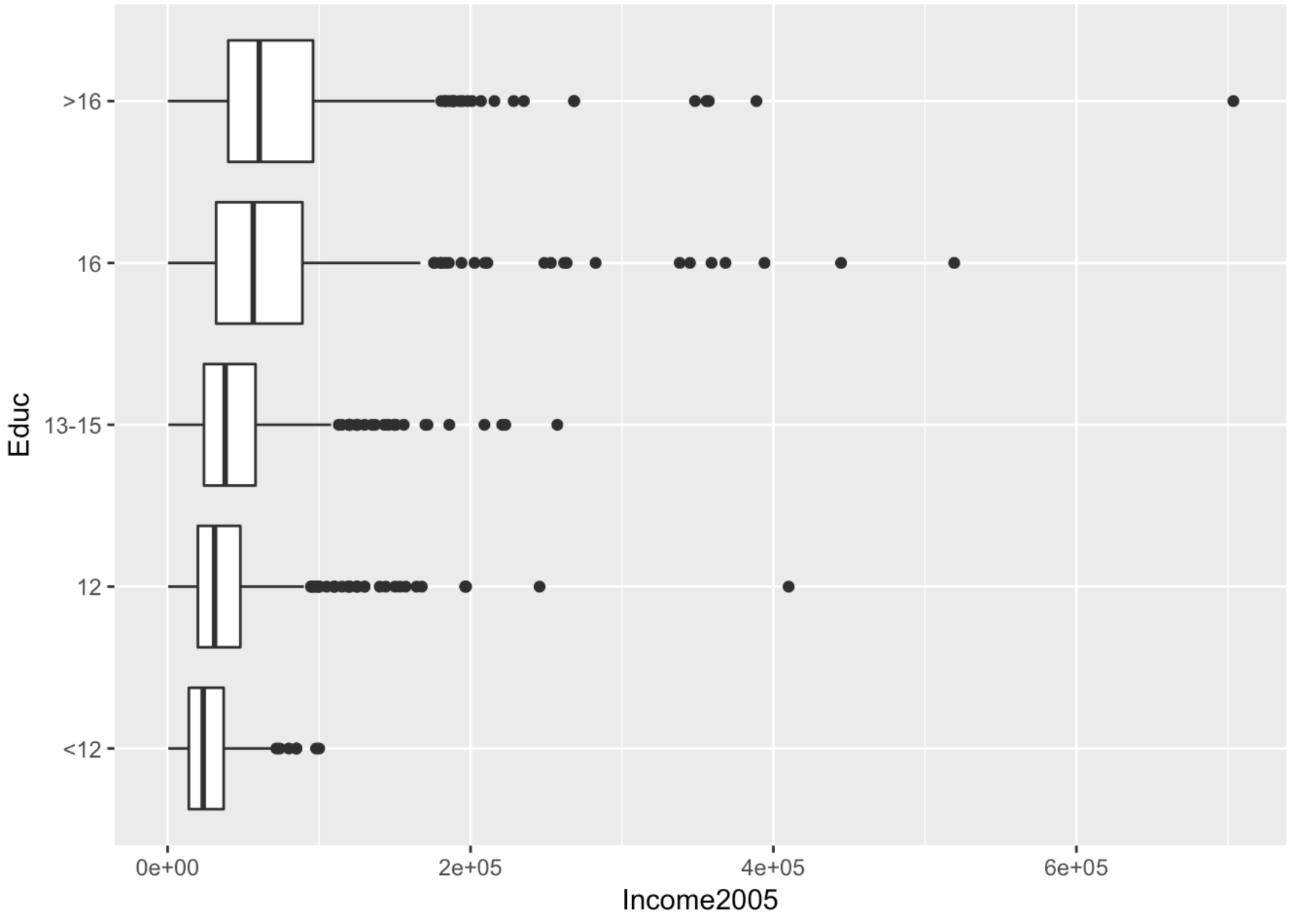
1. Income

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")
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
```



We can make the following predictions from the above boxplot:-

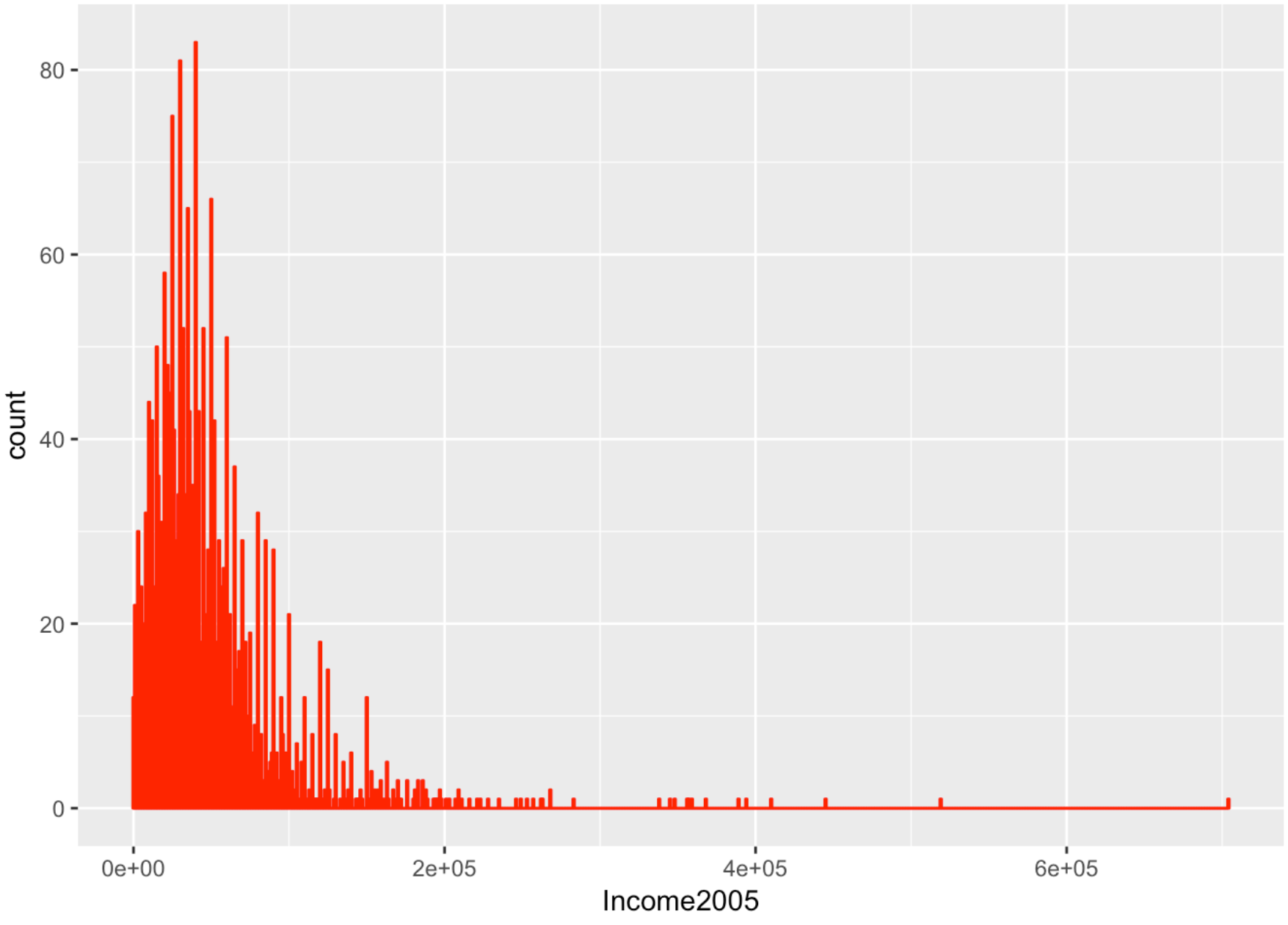
1. We notice that as the Education level increases, the data becomes more right-skewed, i.e most of the data is towards the lower quartiles.
2. The median value of the income increases, with the increase in education level.The quantiles (all: 25%, 50%, 75%) are higher for people with more years of education, with, '>16' bucket being the highest.
3. The magnitude of outliers(data beyond the right whisker) for each education level increases with the level.

4.The variance in the data increases with the increase in the education level.

5.The highest income in the '12' group is ~100,000 whereas the top 25 percentile for the '16' and '>16' groups earn higher than that

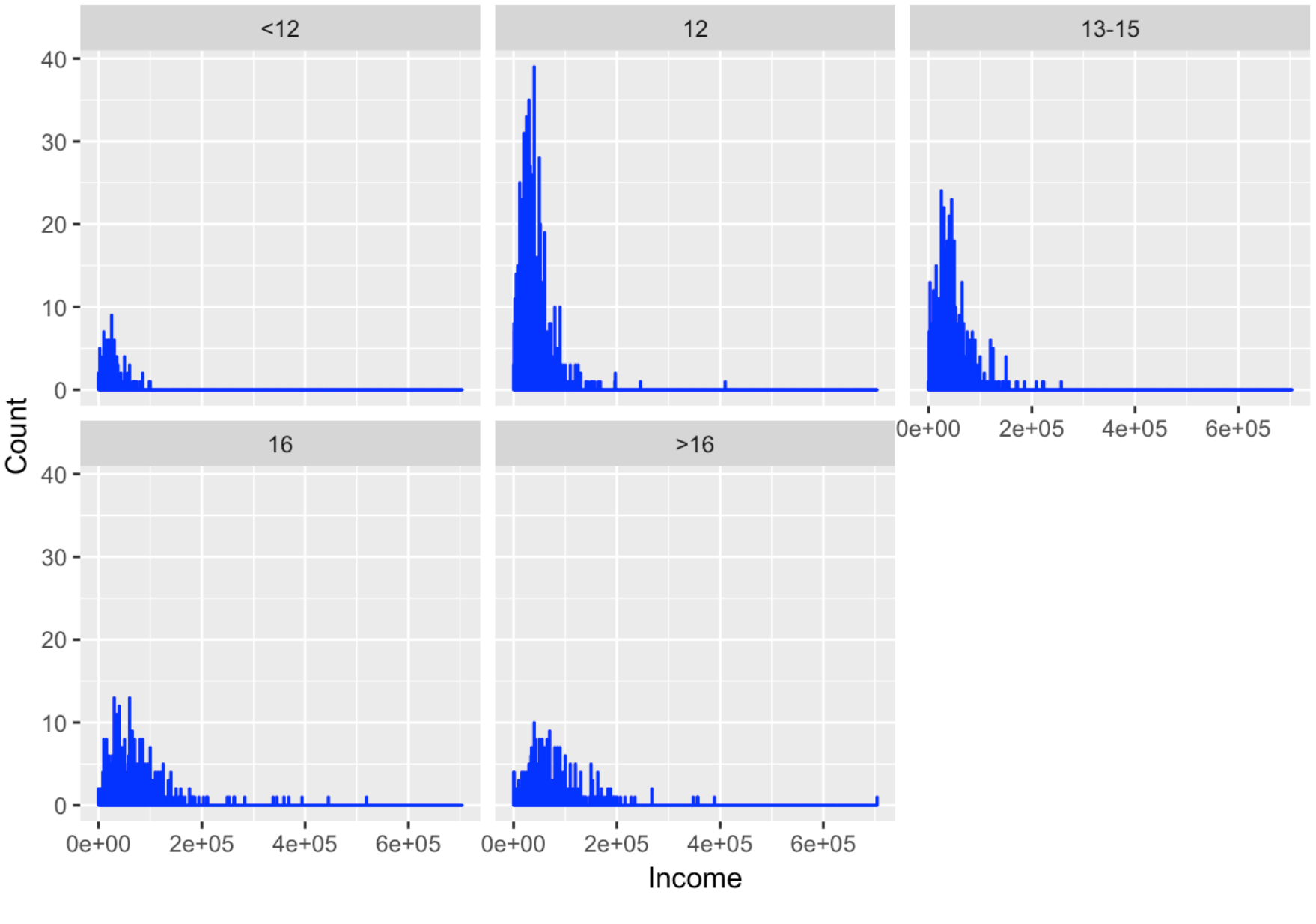
b. Plot a histogram of the `Income2005` variable in the dataset referenced in part a). [3 points]

```
#
ggplot(mydata, aes(x=Income2005)) + geom_histogram(binwidth=1000,colour="red")
```



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

```
ggplot(mydata, aes(x=Income2005)) + geom_histogram(binwidth=1000,colour="blue")+fa
cet_wrap(~Educ)+labs(title="Income vs Years of Education", x="Income", y="Count")
```



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

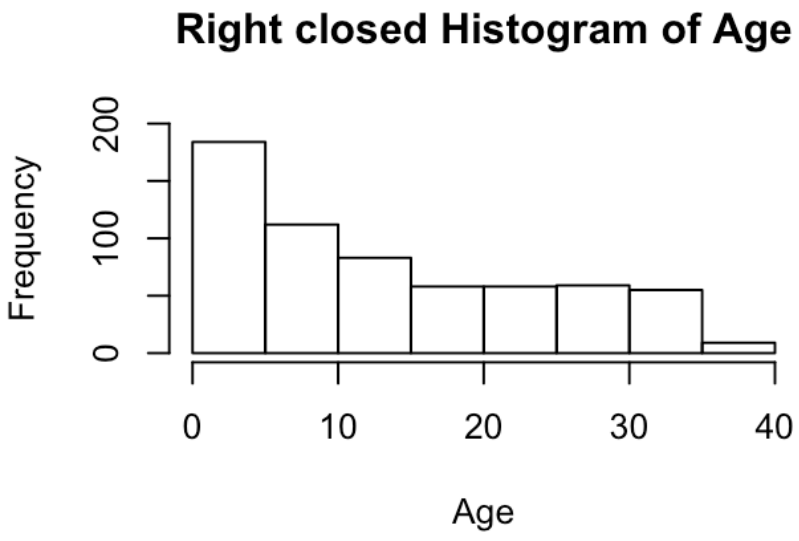
1.The relative number of people in each category is indicated by a histogram, this information cannot be interpreted from the boxplot.

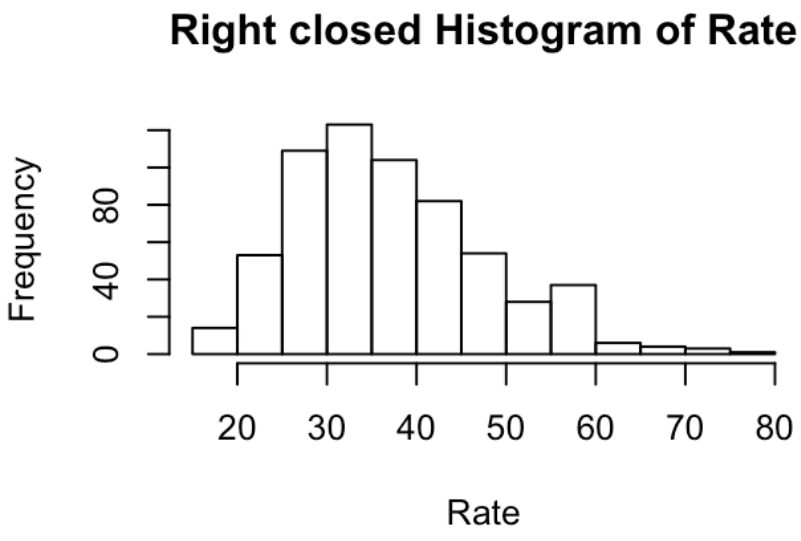
2.The box plot does not give us any information about the mode of the data, whereas the histogram tells us that that the data is unimodal.

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]

```
mydata2<-ex0824
par(mfrow=c(2,2))
hist(mydata2$Age,xlab="Age", main="Right closed Histogram of Age", ylim=c(0,200))
hist(mydata2$Age,right=FALSE,xlab="Age", main="Right open Histogram of Age", ylim=
c(0,200) )
hist(mydata2$Rate,xlab="Rate",  main="Right closed Histogram of Rate")
hist(mydata2$Rate,xlab="Rate",right=FALSE , main="Right open Histogram of Rate")
```



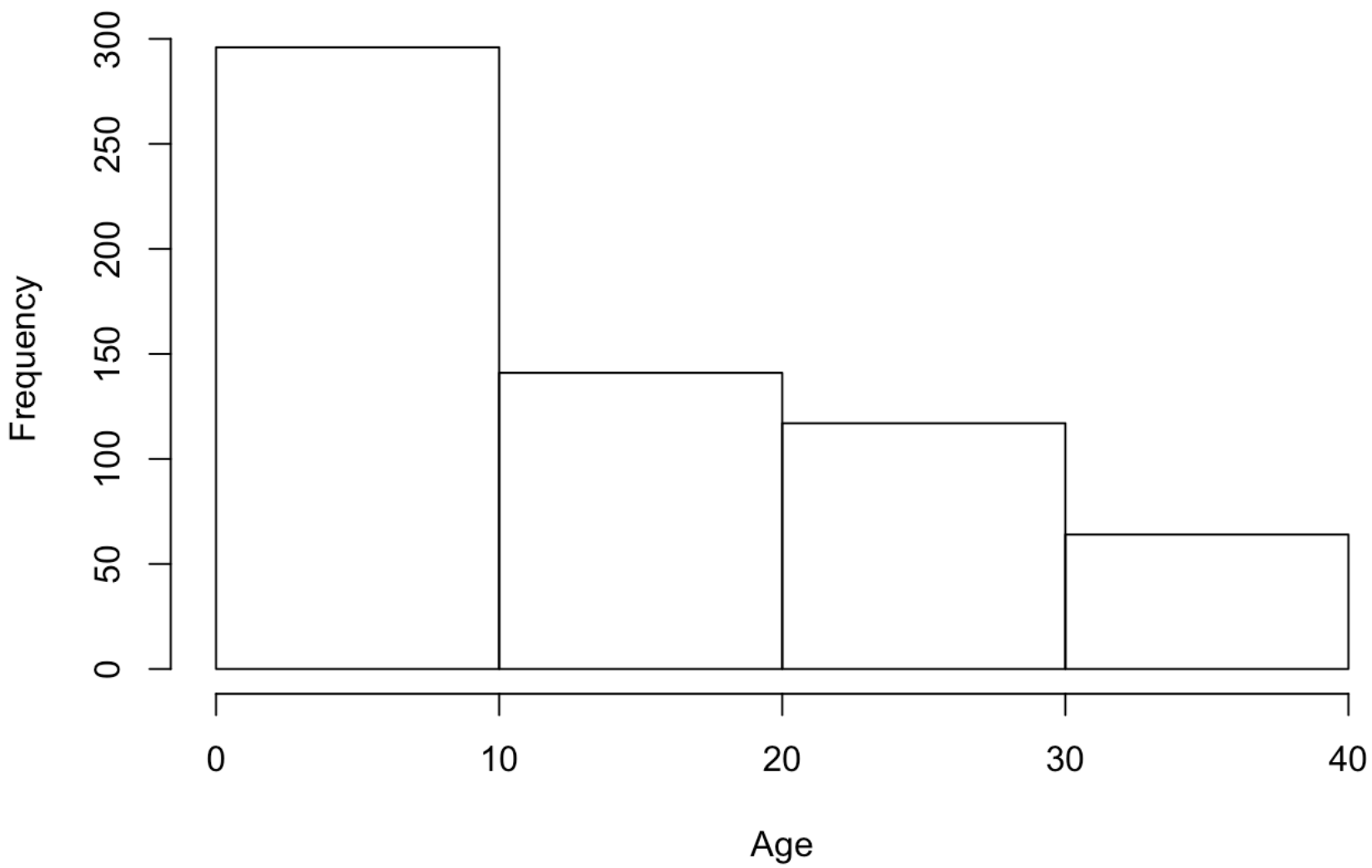


b. For which variable, `Age` or `Rate` , do the two versions differ more? Why? [3 points]

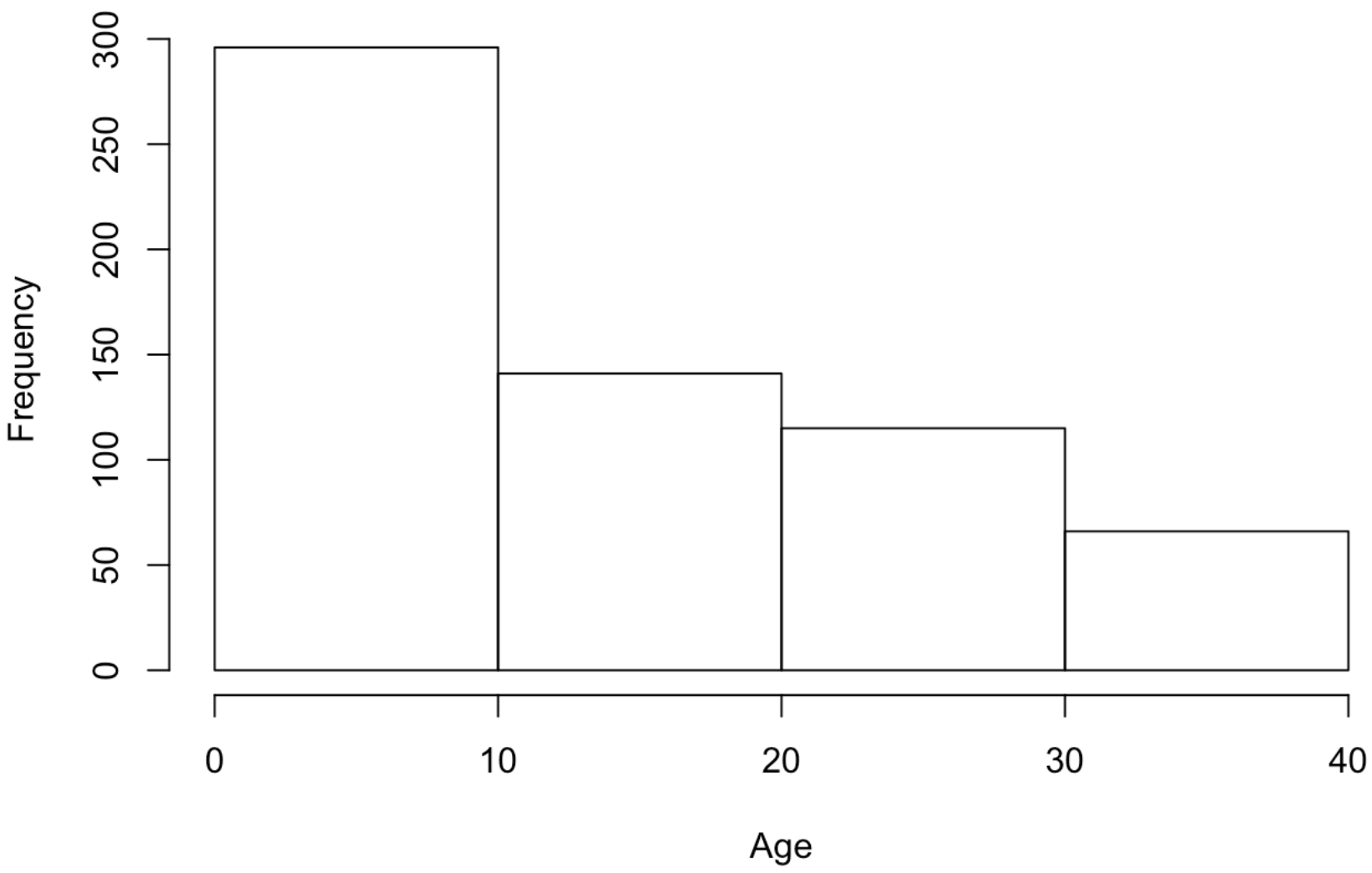
The right closed and right open histograms for the variable `Rate` differ more than the variable `Age` . We can explain this characteristic by noticing that the variable Age would take decimal values(Age in years and months would be rounded off to the nearest year) which are rounded off to the nearest boundary value. Hence, there is no difference between the right closed and right open histograms of the Age variable. On the contrary,rate has whole number values which may fall on the bin boundaries, which is why the right closed and the right open histograms differ significantly.

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

```
mydata2<-ex0824
hist(mydata2$Age,xlab="Age", breaks=4,main="Right closed Histogram of Age")
```



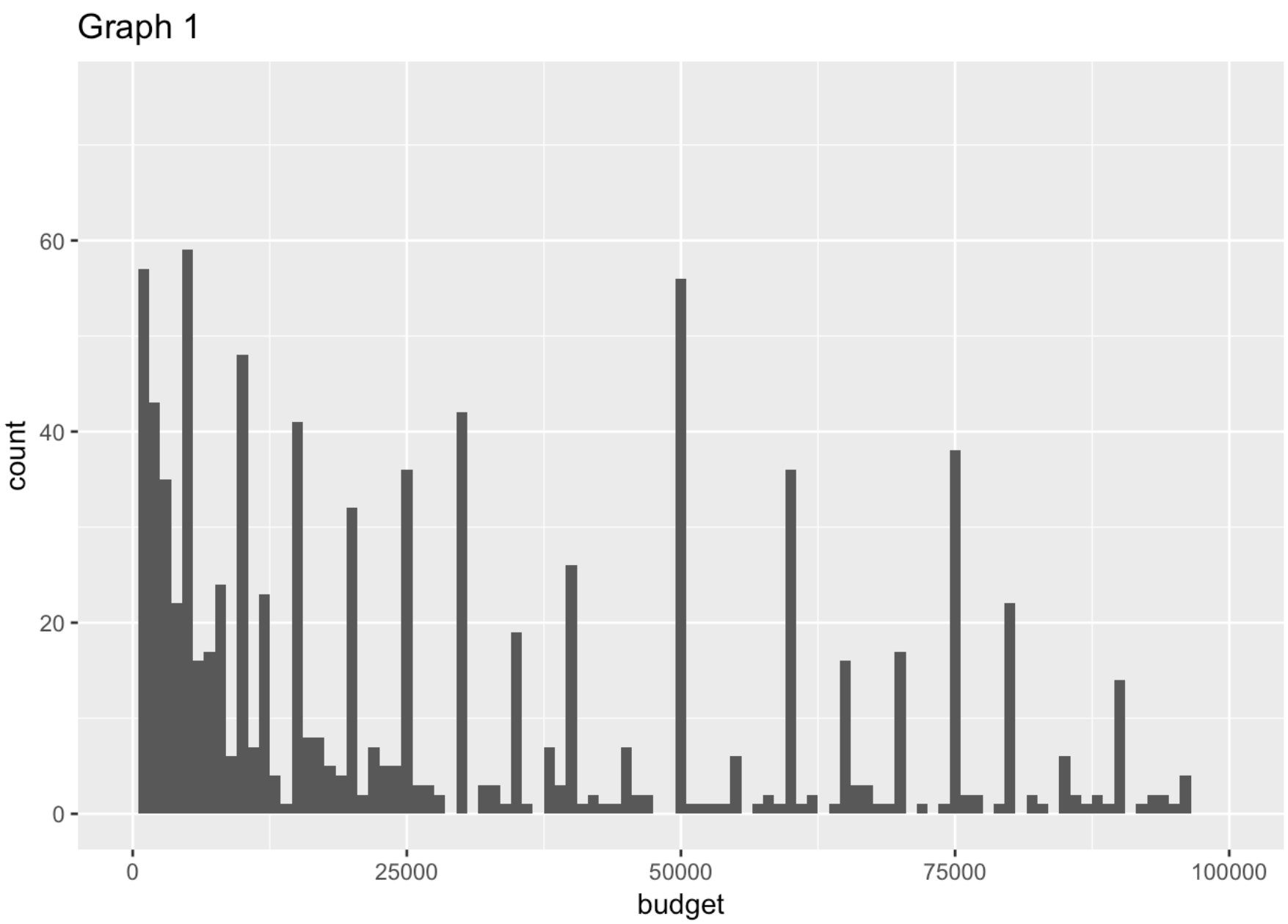

```
hist(mydata2$Age,right=FALSE, breaks=4, xlab="Age", main="Right open Histogram of
Age")
```



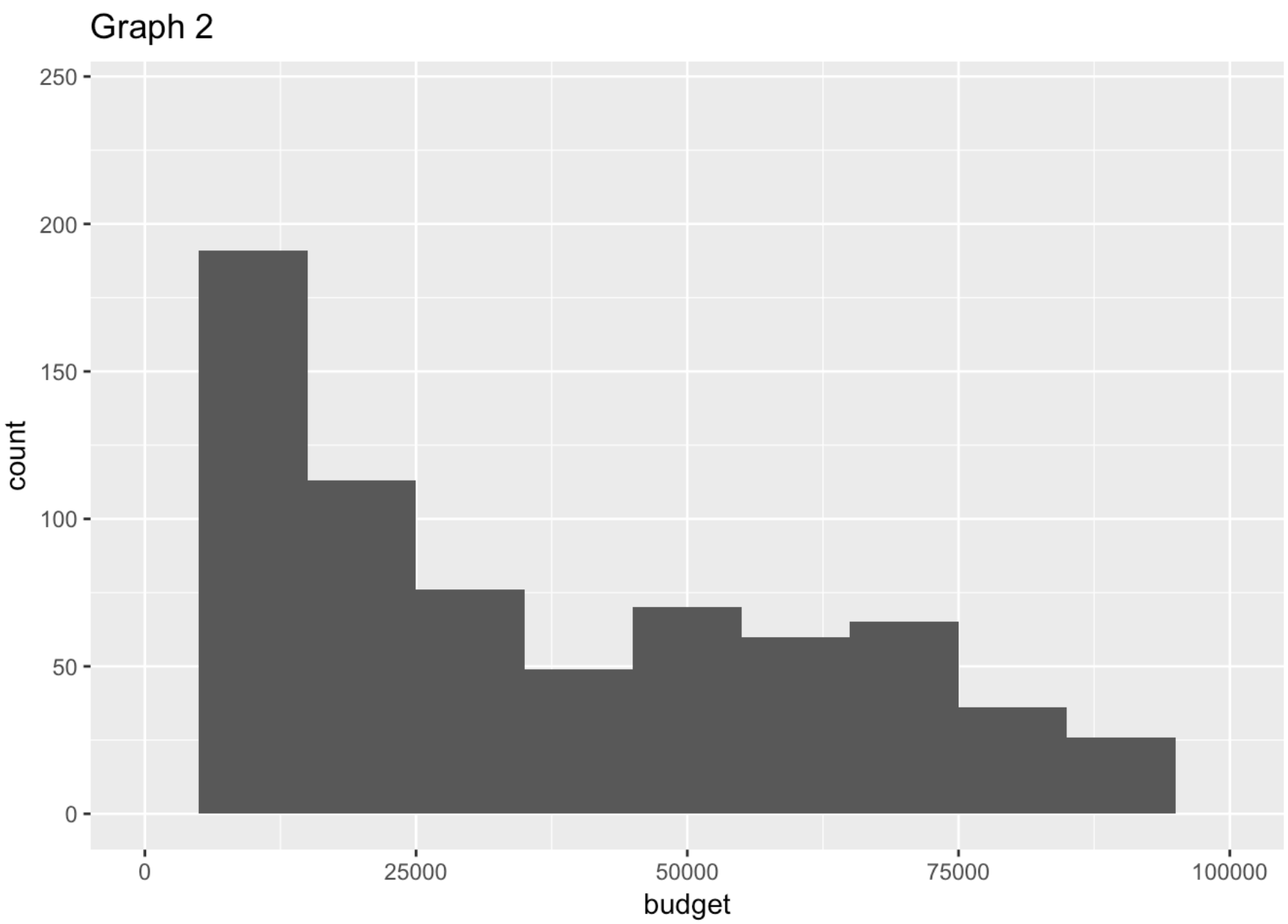
3. Movie budgets

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]

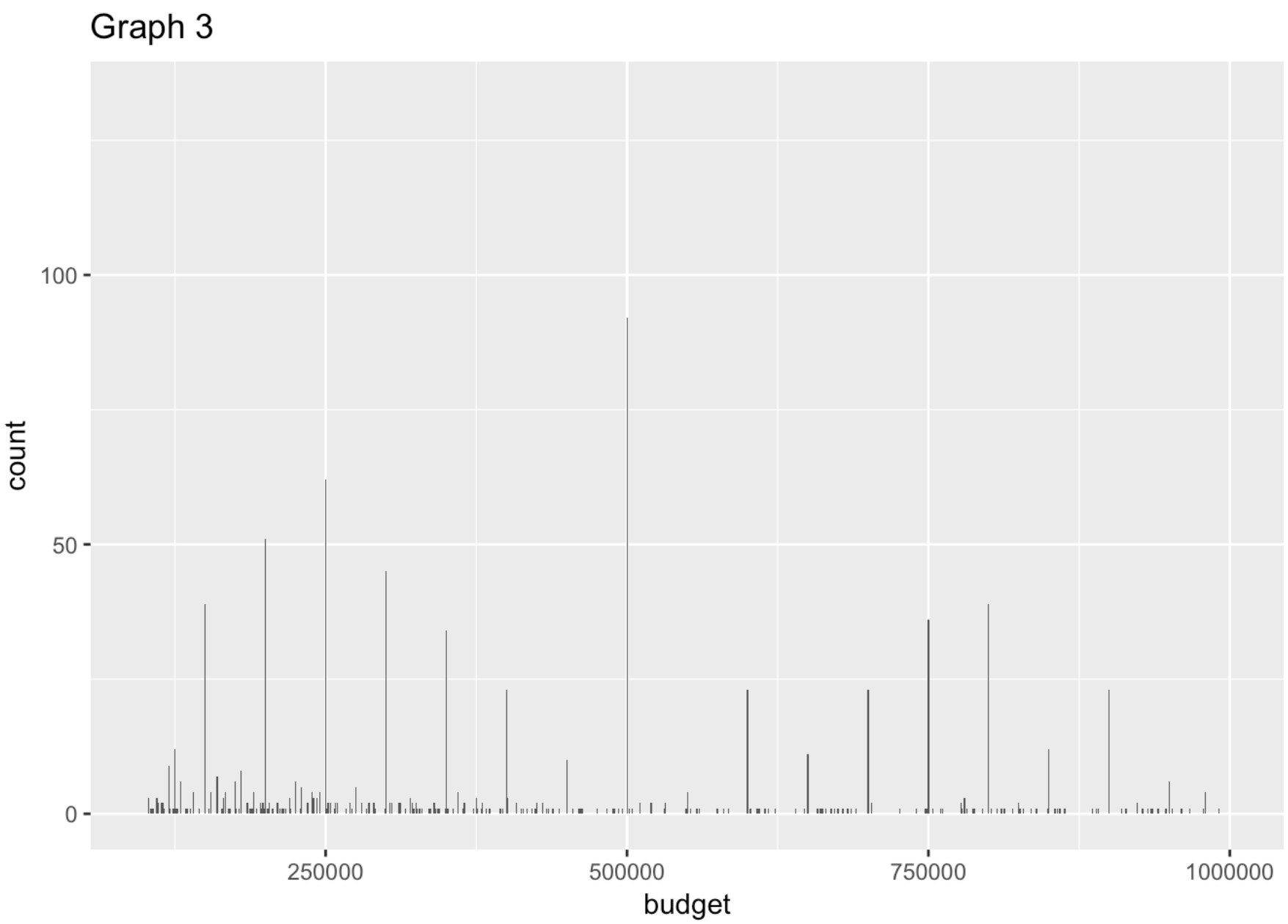
```
#install.packages("ggplot2movies")
#install.packages("ggplot")
library(ggplot2)
library(ggplot2movies)
ggplot(movies,aes(budget))+geom_histogram(binwidth=1000)+xlim(0,100000)+ggtitle(("
Graph 1"))
```



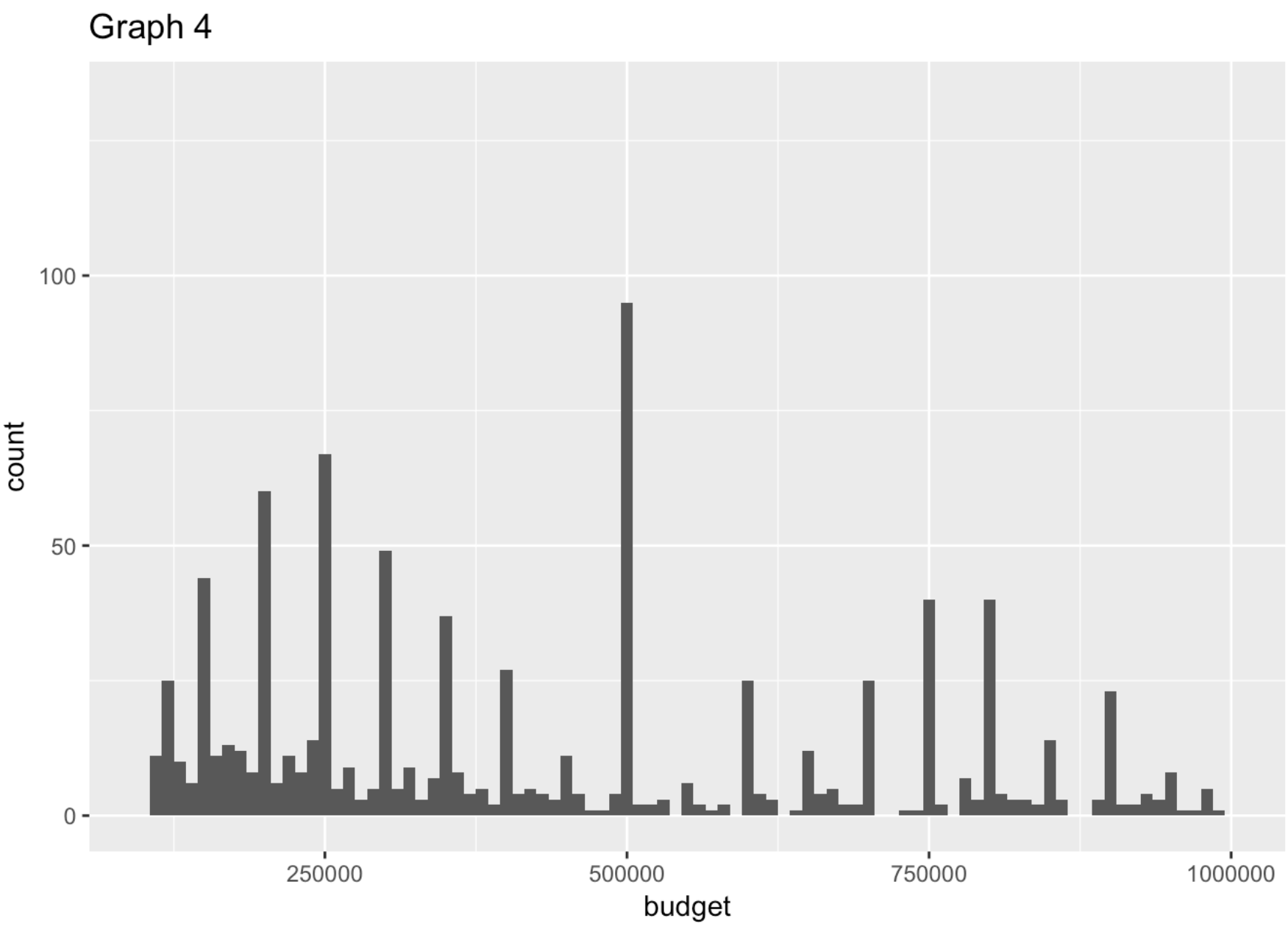

```
ggplot(movies,aes(budget))+geom_histogram(binwidth=10000)+xlim(0,100000)+ggtitle((
"Graph 2"))
```




```
ggplot(movies,aes(budget))+geom_histogram(binwidth=1000)+xlim(100000,1000000)+ggti
tle(("Graph 3"))
```




```
ggplot(movies,aes(budget))+geom_histogram(binwidth=10000)+xlim(100000,1000000)+ggt
itle(("Graph 4"))
```

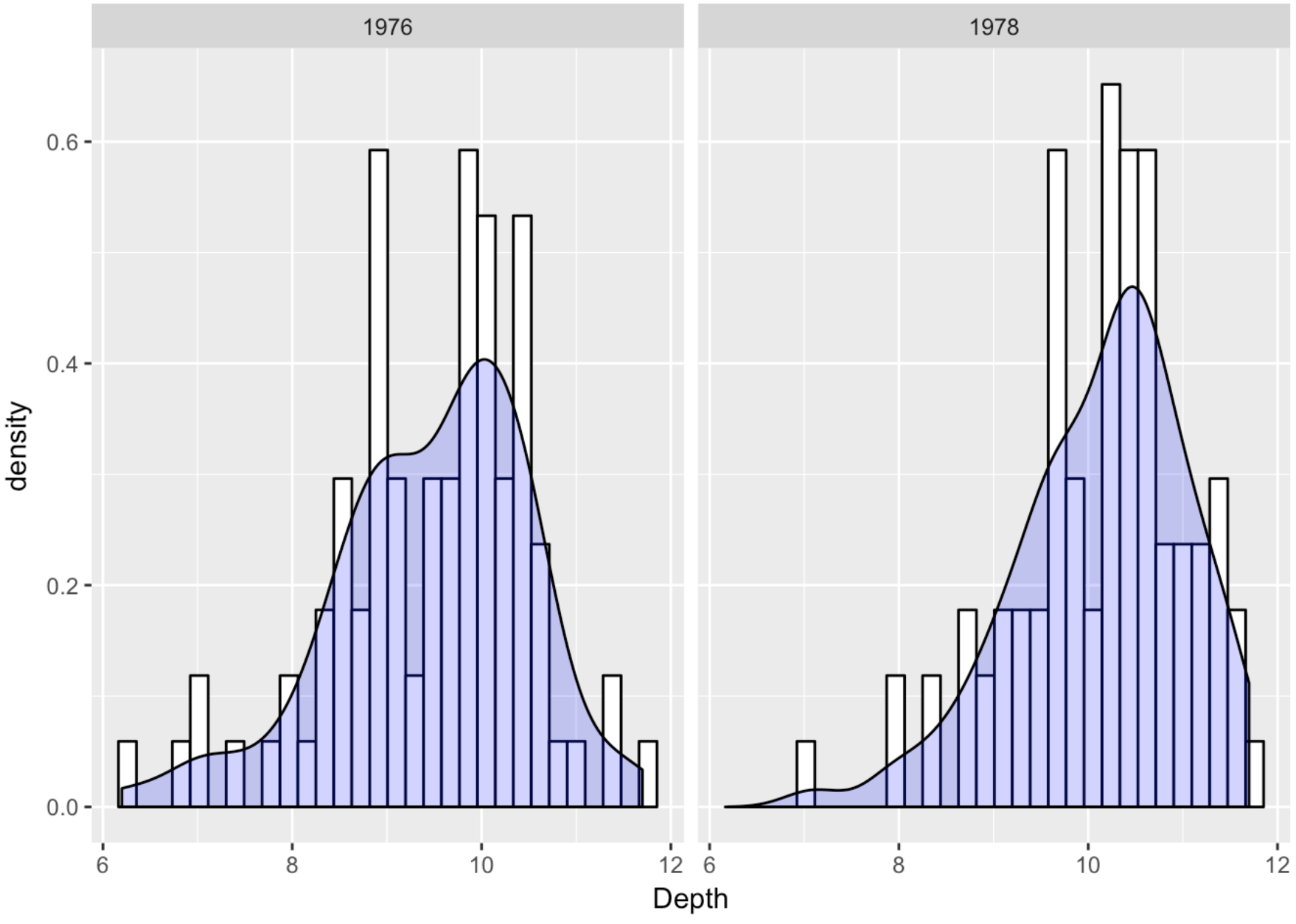


From Graph 1, Graph 3 and Graph 4, we observe that there are several peaks in the data, (like around 5000 for Graph 1 which represents the low budget movies and around 50000 for Graph 3 and 4 which represent the high budget movies) , which signifies that there are several values being rounded off in these ranges.For Graph 2, although we cannot observe the peaks significantly due to the large binwidth,we still can see few peaks in the 25000 range.

4. Finches

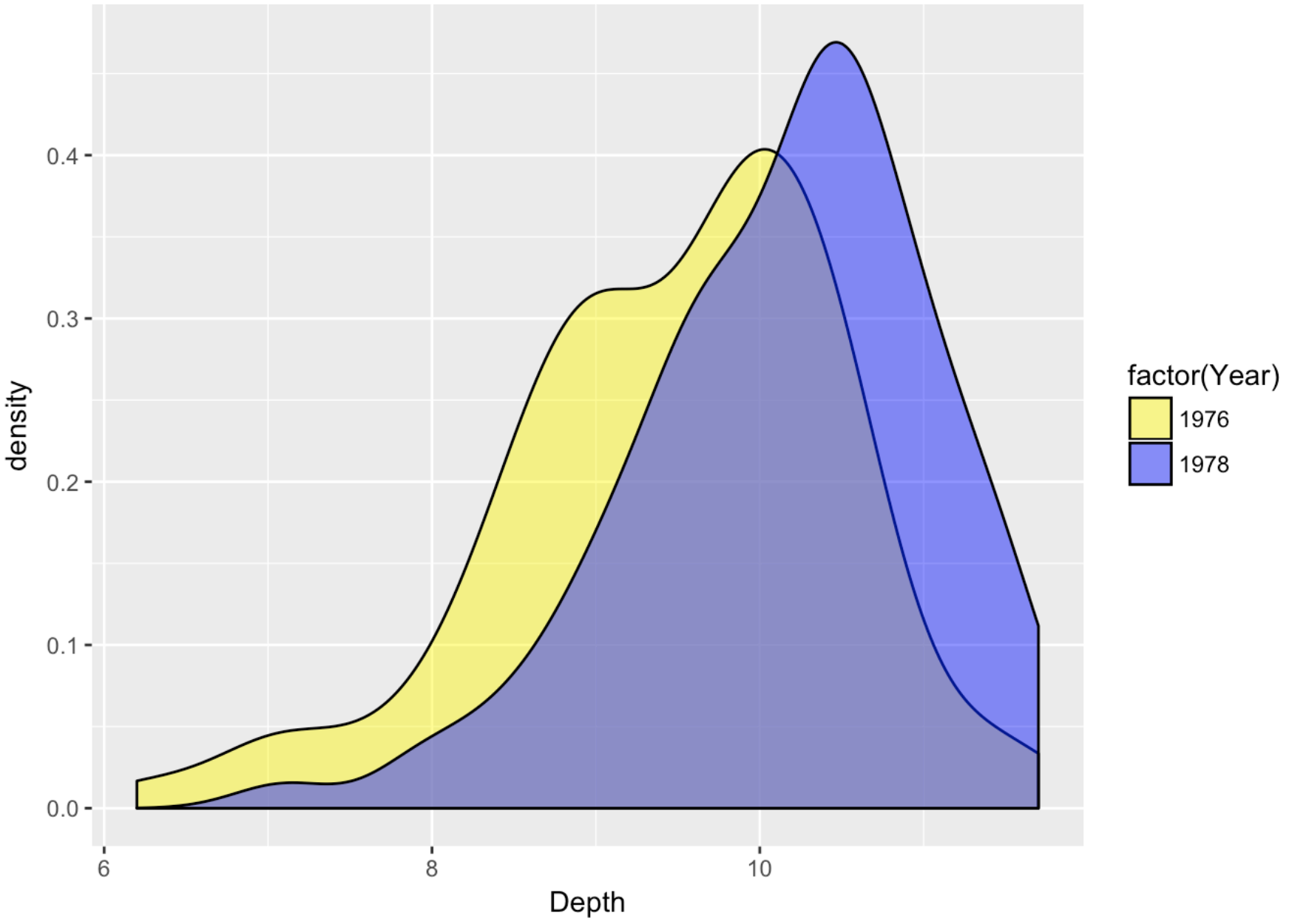
a. Plot separate density histograms of the beak depth of the finches in *case0201* from the **Sleuth3** package, with density curves overlaid as on page 34 of the textbook. (However, do this by facetting on `Year` rather than using `grid.arrange`). [3 points]

```
library(Sleuth3)
ggplot(case0201,aes(x=Depth))+geom_histogram(aes(y=..density..), colour="black", f
ill="white")+geom_density(alpha=.2, fill="blue") + facet_wrap(~Year)
```



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]

```
ggplot(case0201, aes(x= Depth, fill = factor(Year))) +
 geom_density(alpha=.5)+
 scale_fill_manual(values=c("yellow","blue"))
```



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

From the graphs in part a and b, we can infer that there is a difference in the distributions. We observe that the variance of the distribution decreases in the data from 1976 to 1978 resulting the mean depth in 1976 to increase in 1978.

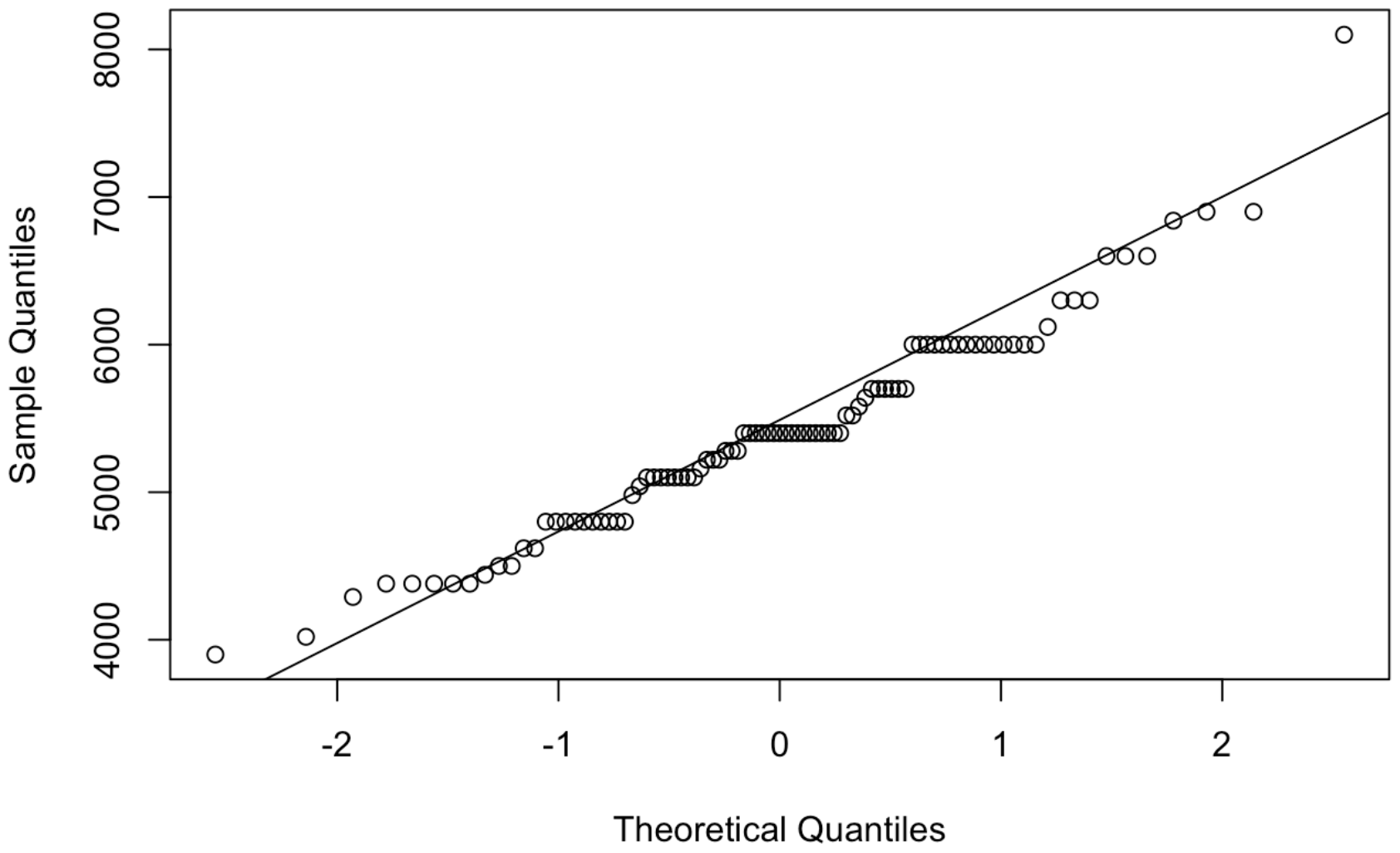
d. What is the cause of the difference according to the information in the help file? [3 points]

According to the information in the help file, in one of those years, 1977, a severe drought caused vegetation to wither, and the only remaining food source was a large, tough seed, which the finches ordinarily ignored.The birds with larger and stronger beaks for opening these tough seeds more likely to survive that year, and did they tend to pass this characteristic to their offspring. Hence, the density of larger beak birds increases after the drought compared to the shorter ones.

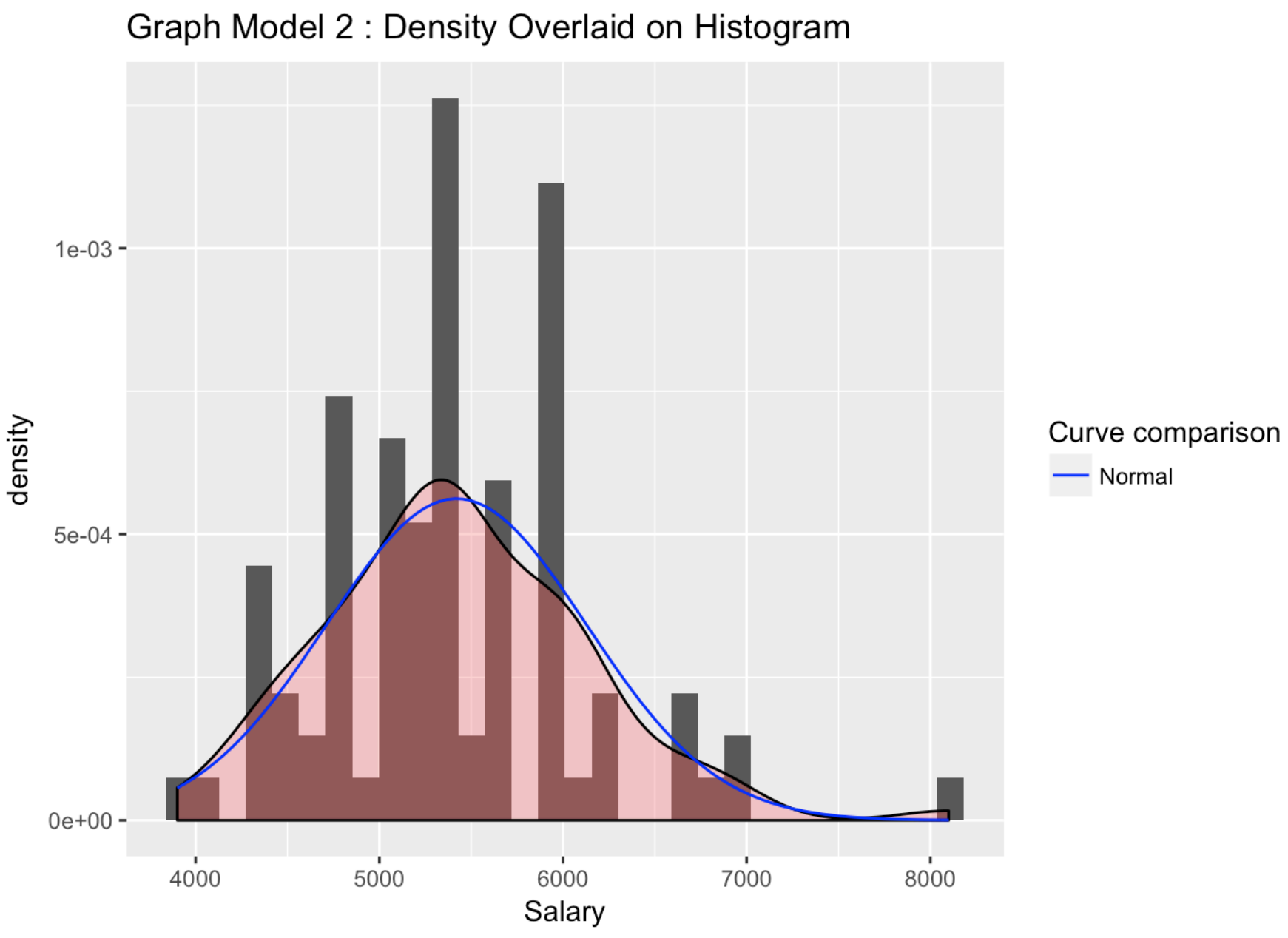
5. Salary

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

```
qqnorm(case0102$Salary,main="Graph Model 1 : Q-Q plot of case0102 Salary")
qqline(case0102$Salary,distribution = qnorm)
```




```
ggplot(case0102,aes(Salary))+geom_histogram(aes(y=..density..))+geom_density(alpha
=0.3,fill="#FF6666")+stat_function(fun = dnorm, args = with(case0102,c(mean(case01
02$Salary),sd(case0102$Salary))), aes(color = "Normal"))+
  scale_color_manual("Curve comparison", values = c("blue"))+ggtitle("Graph Model
2 : Density Overlaid on Histogram")
```



According to the graphs plotted, the data seems to be closely normally distributed.