STAT 327 Homework 3

We'll grade your homework by

- opening your "hw3.Rmd" file in RStudio
- clicking "Knit HTML"
- reading the HTML output
- reading your "hw3.Rmd"

You should write R code anywhere you see an empty R code chunk. You should write English text anywhere you see "…"; please surround it with doubled asterisks (`**...**`) so that it will show up as boldface and be easy for us to find.

Include reasonable labels (titles, axis labels, legends, etc.) with each of your graphs.

Name: Naiqing Cai

Email: ncai5@wisc.edu (mailto:ncai5@wisc.edu)

We'll use data on housing values in suburbs of Boston. They are in an R package called "MASS." (An R package is a collection of code, data, and documentation. "MASS" refers to the book "Modern Applied Statistics with S." R developed from the earlier language, S.) The MASS package comes with the default R installation, so it's already on your computer. However, it's not loaded into your R session by default. So we'll load it via the `require()` command (there's nothing for you to do here):

```
require("MASS")
```

```
## Loading required package: MASS
```

Run `?Boston` (outside this R Markdown document) to read the help page for the `Boston` data frame.

Convert the `chas` variable to a factor with labels "off" and "on" (referring to the Charles river).

```
Boston$chas = factor(Boston$chas, levels = c(0, 1), labels = c("off", "on"))
```

How many rows are in the Boston data frame? How many columns?

```
b=Boston
n.rows=dim(b)[1]
cat(sep = "", "The number of rows are ", n.rows, "\n")
```

```
## The number of rows are 506
```

```
n.cols=length(b)
cat(sep = "", "The number of columns are ", n.cols, "\n")
```

```
## The number of columns are 14
```

What does a row represent?

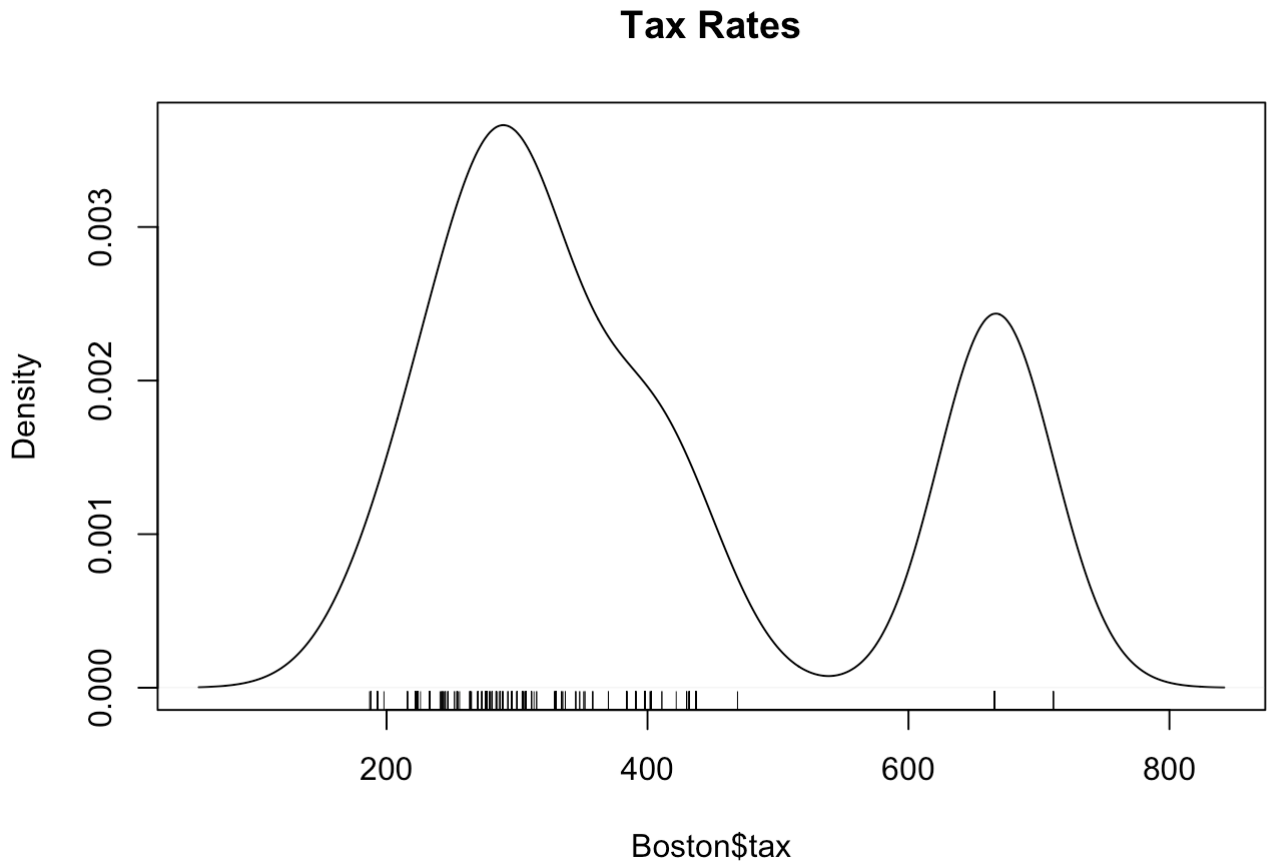
Each row means a unique suburb in the Boston, containing several statistics of that suburb.

What does a column represent?

Each column represents a different statistic that is being measured for each suburb. A column is a set of all observations of a particular statistic.

Make a density plot (with rug) of tax rates.

```
plot(density(Boston$tax),xlab='Boston$tax',main = "Tax Rates")
rug(Boston$tax)
```

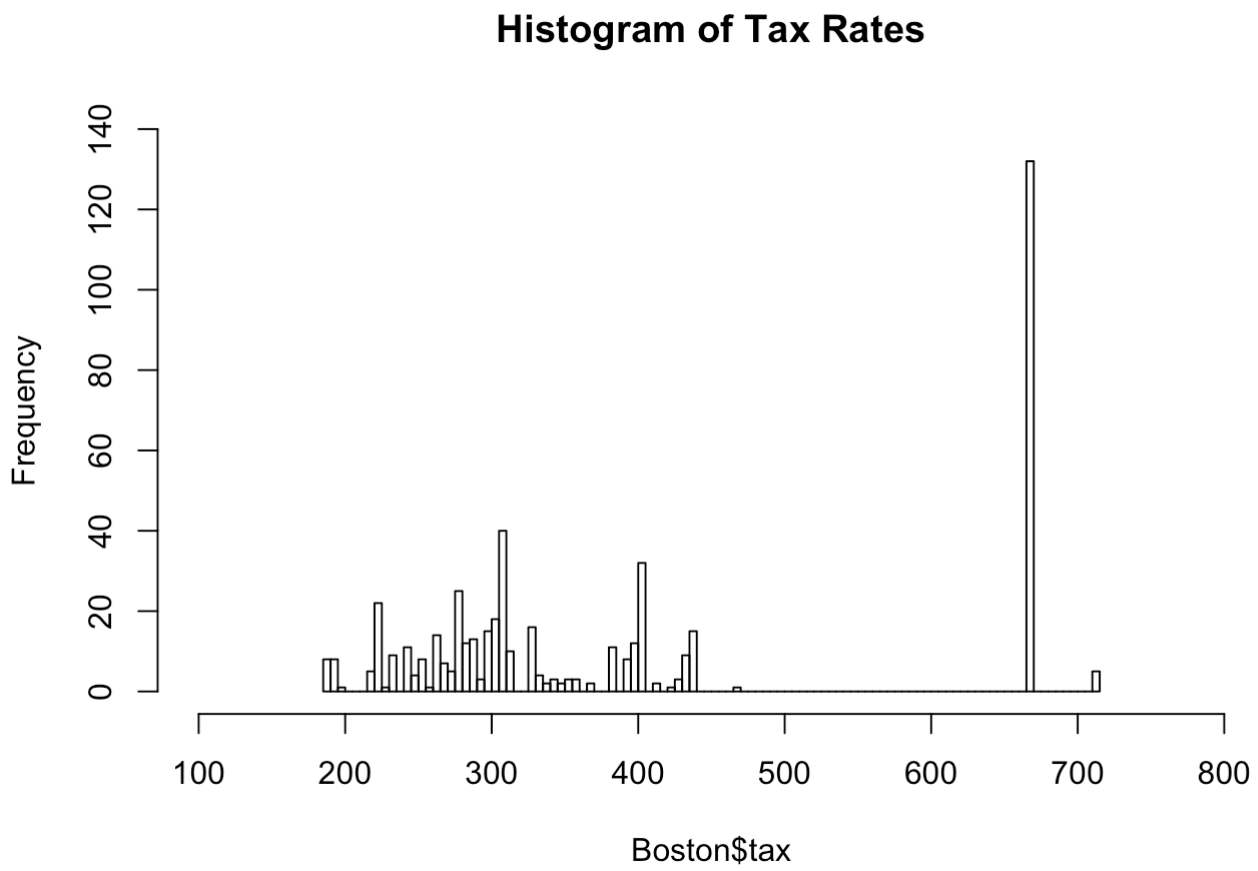


Describe the shape of the distribution of tax rates.

It is Bimodal around approximately 300 and 700.

Note that the distribution shape doesn't make sense in light of the rug representation of the data. Make a histogram of the tax rates.

```
hist(Boston$tax,breaks=100,freq=TRUE,main = "Histogram of Tax Rates",ylim = c(0, 140
), xlim = c(100, 800))
```



Why is the second peak of the density plot so large? In what way is the rug representation of the data inadequate? Write a line or two of code to figure it out, and then explain it.

```
max = length(which(Boston$tax == max(Boston$tax)))
cat(sep = "", "There occurs ", max, " times the maxium value ",max(Boston$tax),
    " in this histogram.", "\n")
```

```
## There occurs 5 times the maxium value 711 in this histogram.
```

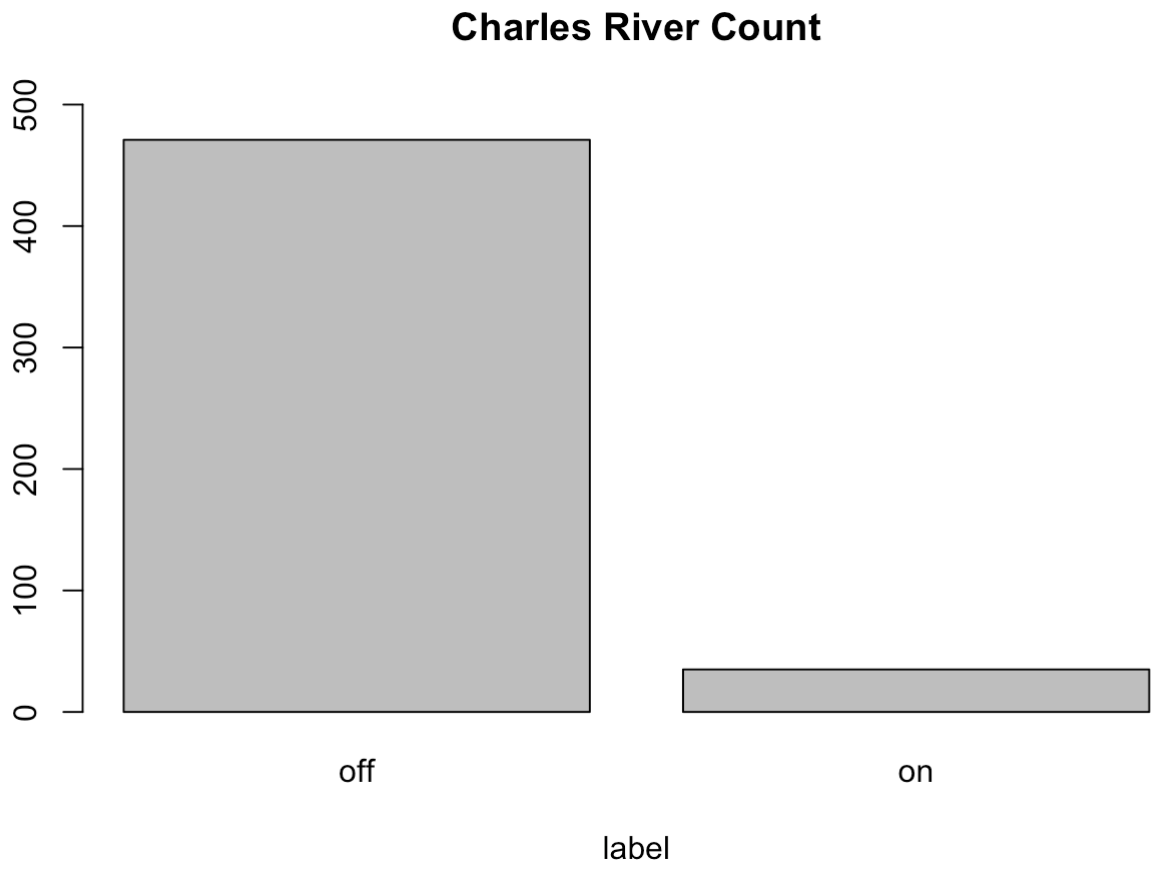
```
second = sort(Boston$tax, TRUE)[max + 1]
secondcount = length(which(Boston$tax == second))
cat(sep = "", "There occurs ", secondcount, " times the second highest value ",
    second, " in this histogram.", "\n")
```

```
## There occurs 132 times the second highest value 666 in this histogram.
```

Explanation: At the second peak there is a very high frequency of places with a tax rate of 666 and 711, which creates such a large peak. However, on the rug, this is represented as only one dark line, which is misleading the result and that's why there would be a peak there.

Make a barplot of "chas".

```
counts = table(Boston$chas)
barplot(counts, main = "Charles River Count", xlab="label",ylim = c(0,500))
```



How many neighborhoods are on the Charles river?

```
neighborhoods = length(which(Boston$chas == "on"))
cat(sep = "", "There are ", neighborhoods, " neighborhoods on the Charles river.",
"\n")
```

```
## There are 35 neighborhoods on the Charles river.
```

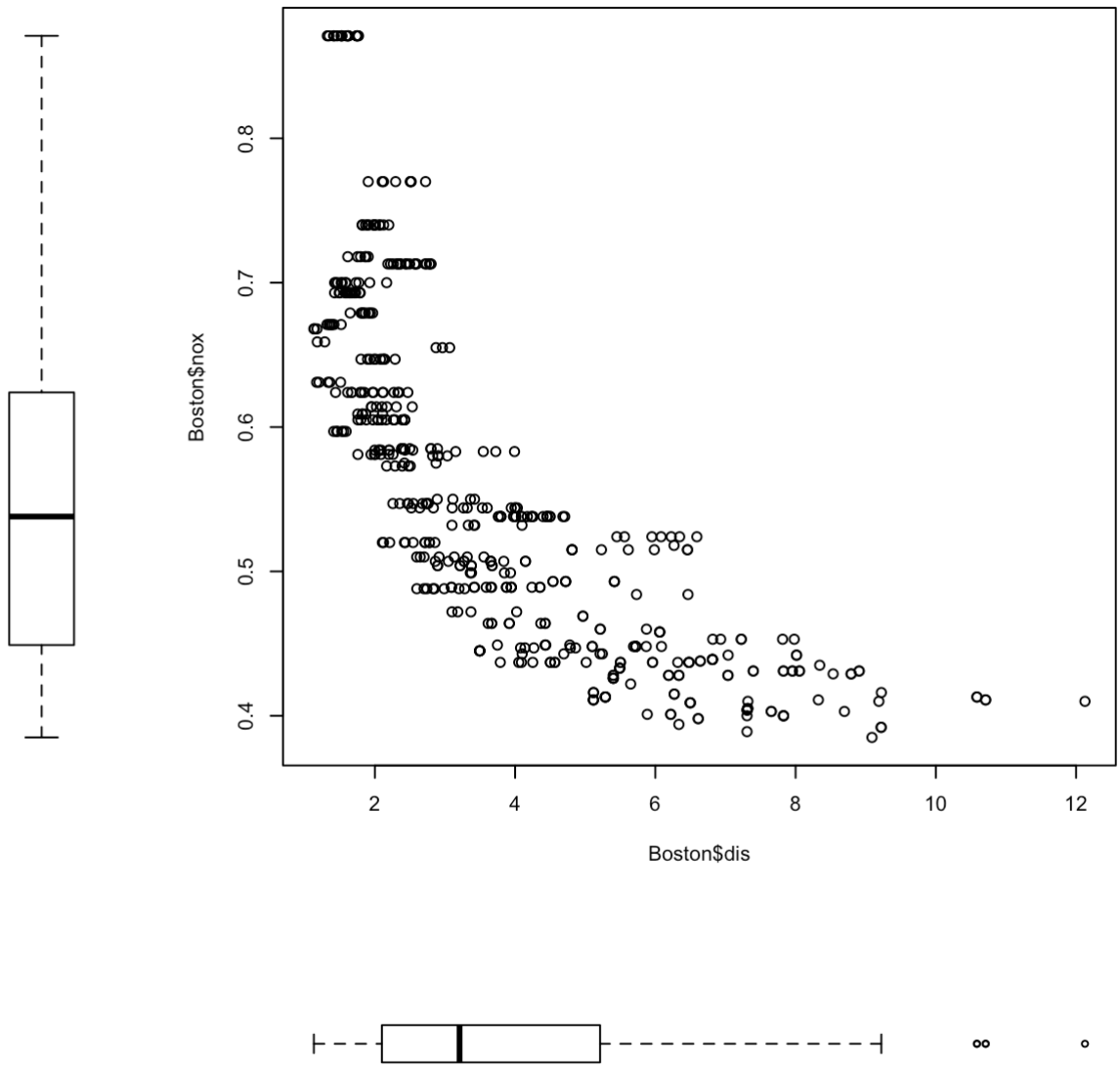
Make a single graph consisting of three plots:

- a scatterplot of “nox” on the y-axis vs. “dis” on the x-axis
- a (vertical) boxplot of “nox” left of the scatterplot’s y-axis
- a (horizontal) boxplot of “dis” below the scatterplot’s x-axis

Hint: use `layout()` with a 4x4 matrix, using the top-right 3x3 corner for the scatterplot, leaving the bottom-left 1x1 corner blank, and using the other parts for the boxplots.

(An optional challenge, worth 0 extra credit points: remove the axis and plot border from each boxplot.)

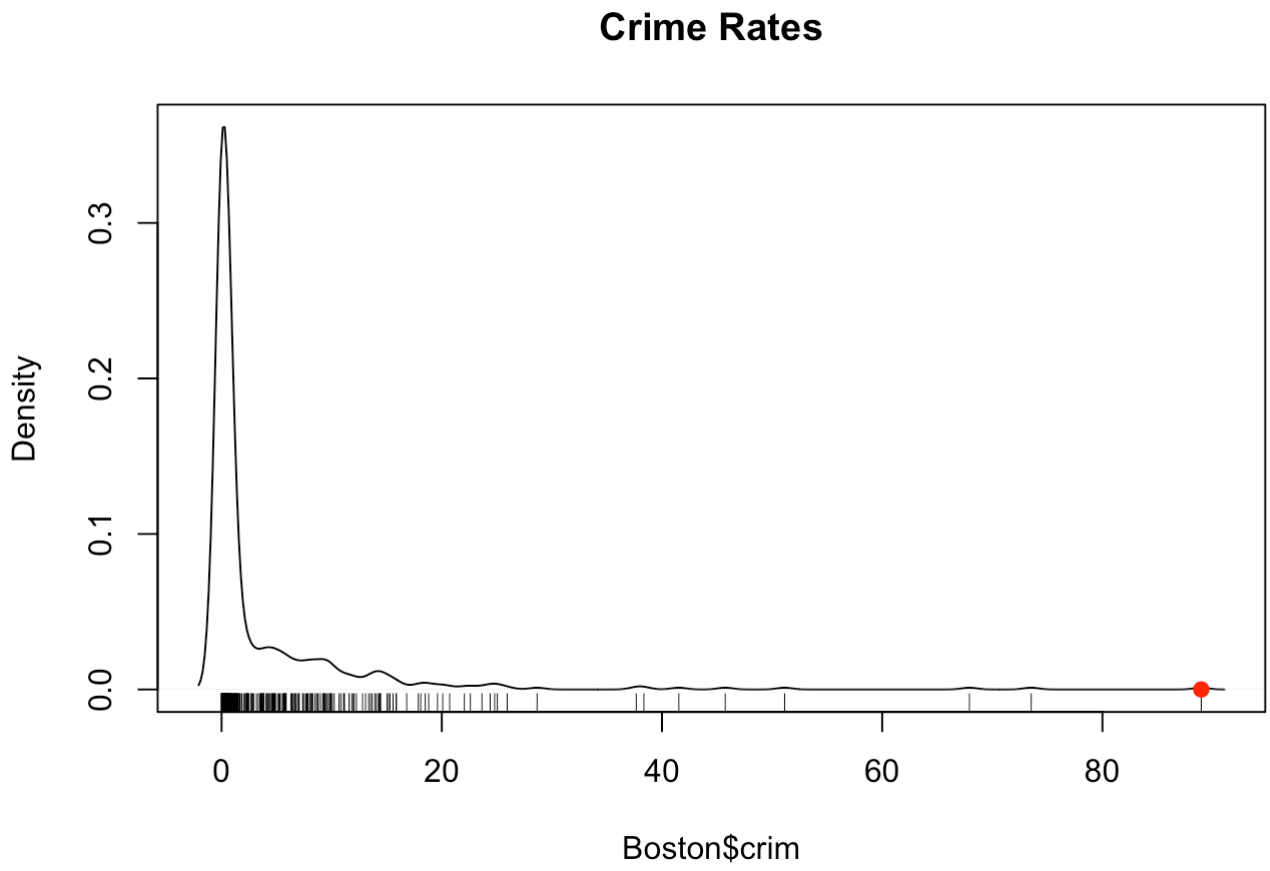
```
layout(matrix(data=c(1, 3, 3, 3, 1, 3, 3, 3, 1, 3, 3, 3, 0, 2, 2, 2), nrow=4, ncol=4,
 byrow=TRUE))
boxplot(x=Boston$nox,axes=FALSE)
boxplot(x=Boston$dis,axes=FALSE,horizontal = TRUE)
plot(Boston$dis,Boston$nox)
```



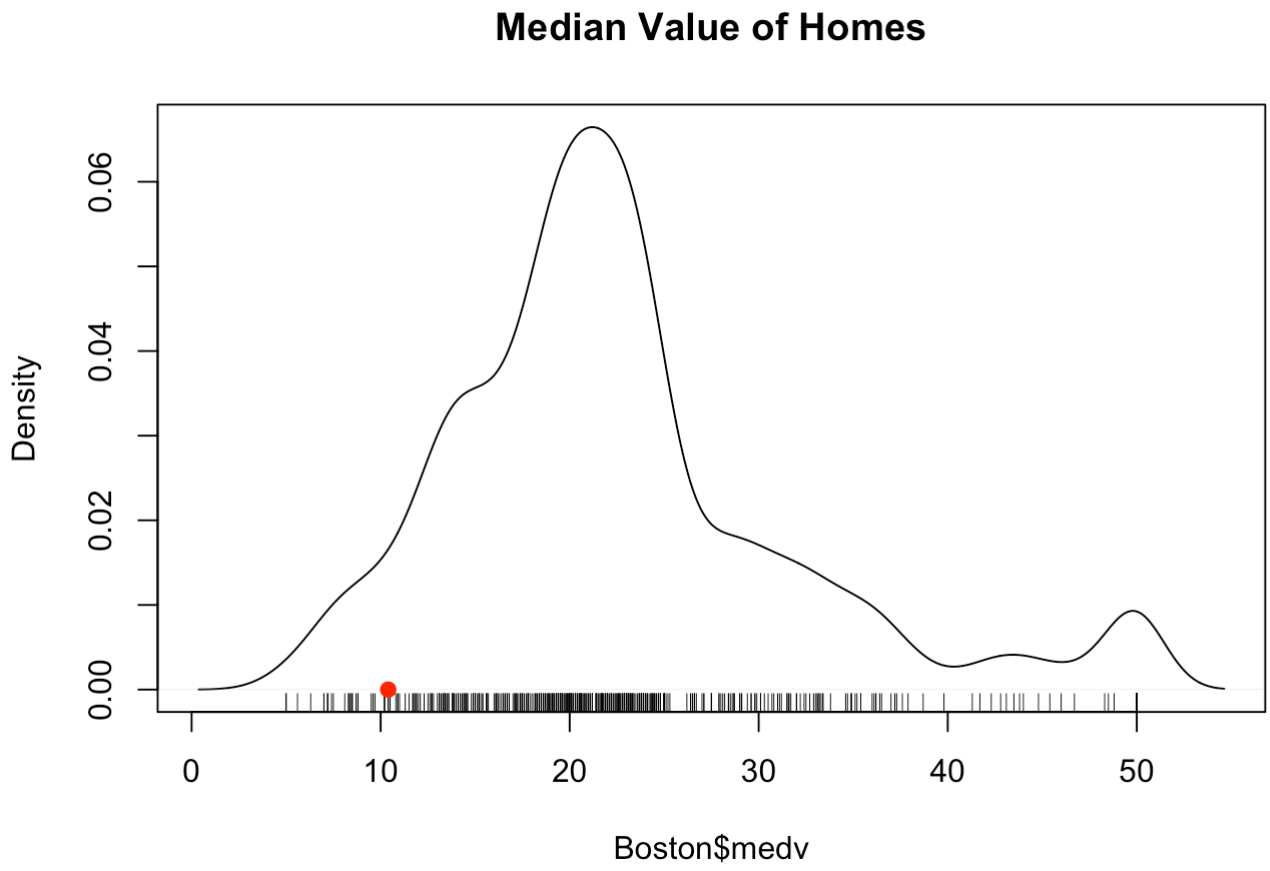
Look into the highest-crime neighborhood by making a single graph of one column of three rows:

- Find the row number, r, of the neighborhood with the highest “crim”.
- Make a density plot of “crim”. Include a rug to show the data.
- Add a red circle at (x, y) = (max crime rate, 0) to make this maximum crime rate stand out.
- Make a density plot with rug of “medv”, adding a red circle at (x, y) = (medv[r], 0) to see what medv corresponds to the highest crime rate.
- Repeat the last step for “ptratio”.

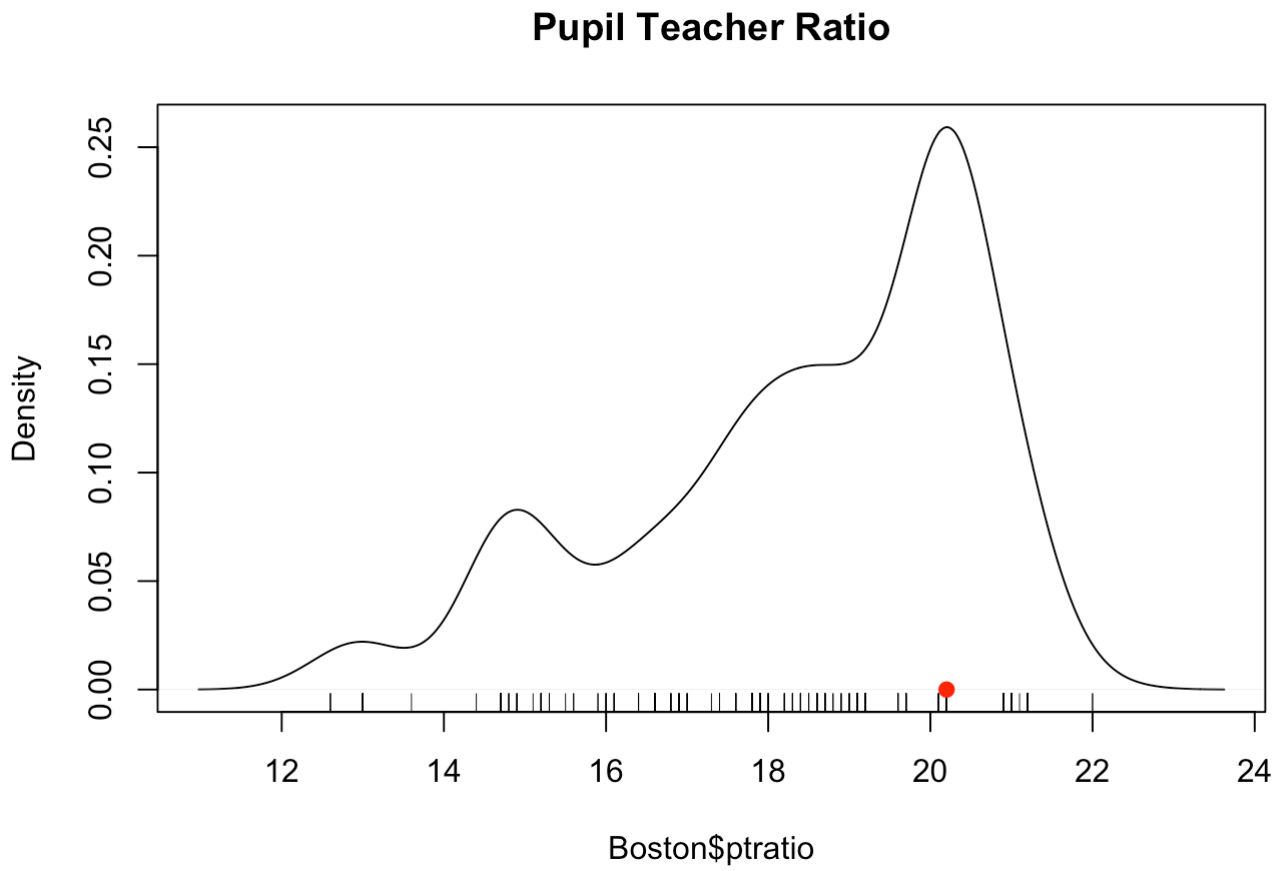
```
r=which(Boston$crim==max(Boston$crim), arr.ind=TRUE)
plot(density(Boston$crim),xlab='Boston$crim',main = "Crime Rates")
rug(Boston$crim)
points(max(Boston$crim),0,col='red',pch = 19)
```




```
plot(density(Boston$medv),xlab='Boston$medv',main = "Median Value of Homes")
rug(Boston$medv)
points(Boston$medv[r], 0,col='red',pch = 19)
```




```
plot(density(Boston$ptratio),xlab='Boston$ptratio',main = "Pupil Teacher Ratio")
rug(Boston$ptratio)
points(Boston$ptratio[r], 0,col='red',pch = 19)
```



What do you notice about the ptratio and medv for the highest-crime neighborhood?

Neighborhood which Pupil Teacher Ratio is about 20, which has the highest density, has the highest-crime. Neighborhood which Median Value of Homes about 10 has the highest-crime.