Lab 1: Exploratory Data Analysis

This Bitter Earth

Paul Nguyen

Exercise 1

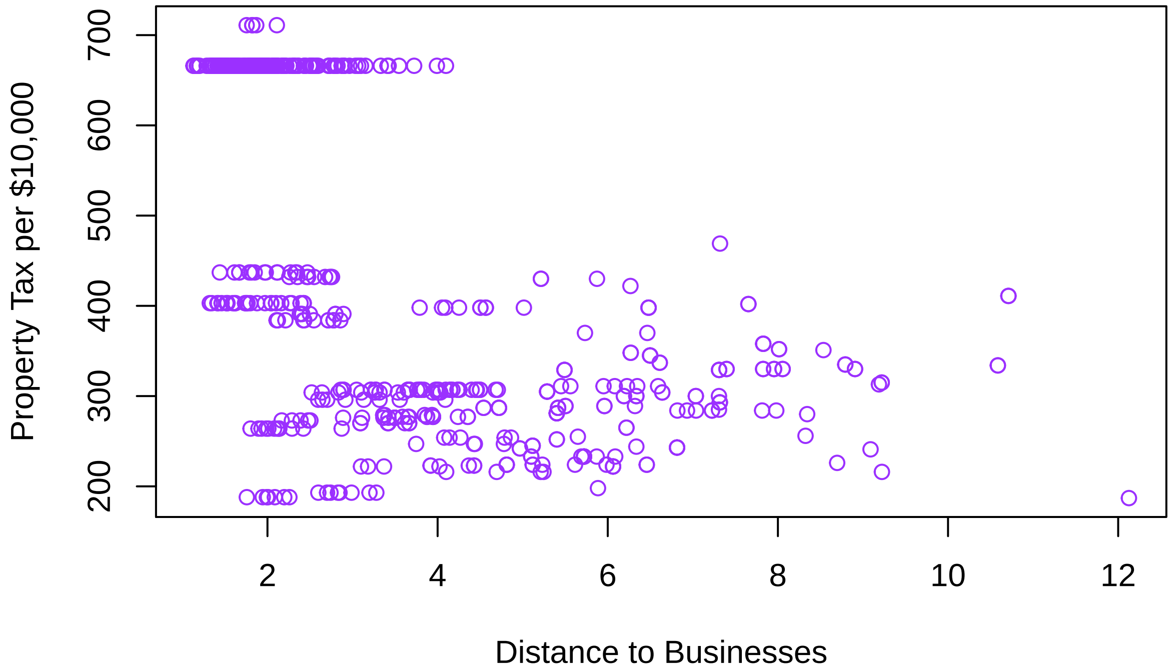
```
library(MASS)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

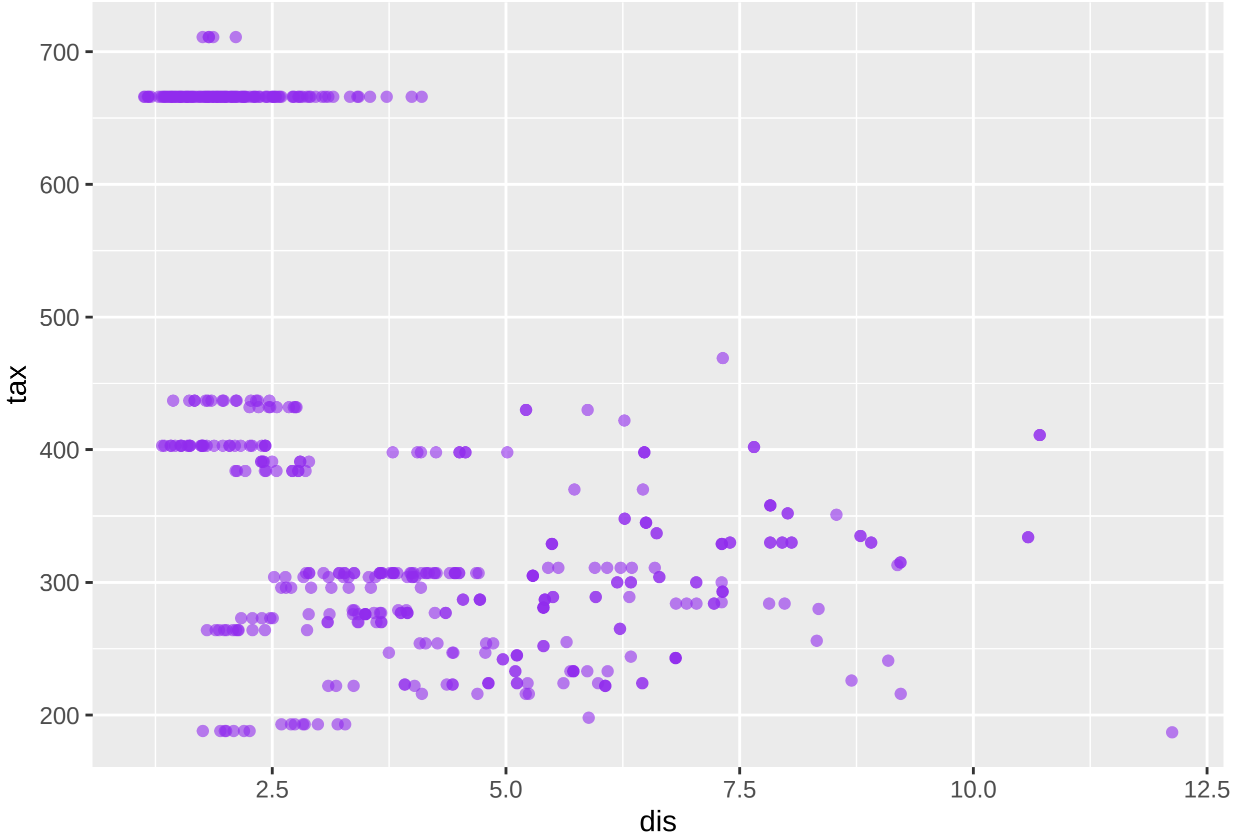
There are 506 rows and 14 columns in this dataset. Each row represents one town, and each column represents a different variable.

Exercise 2

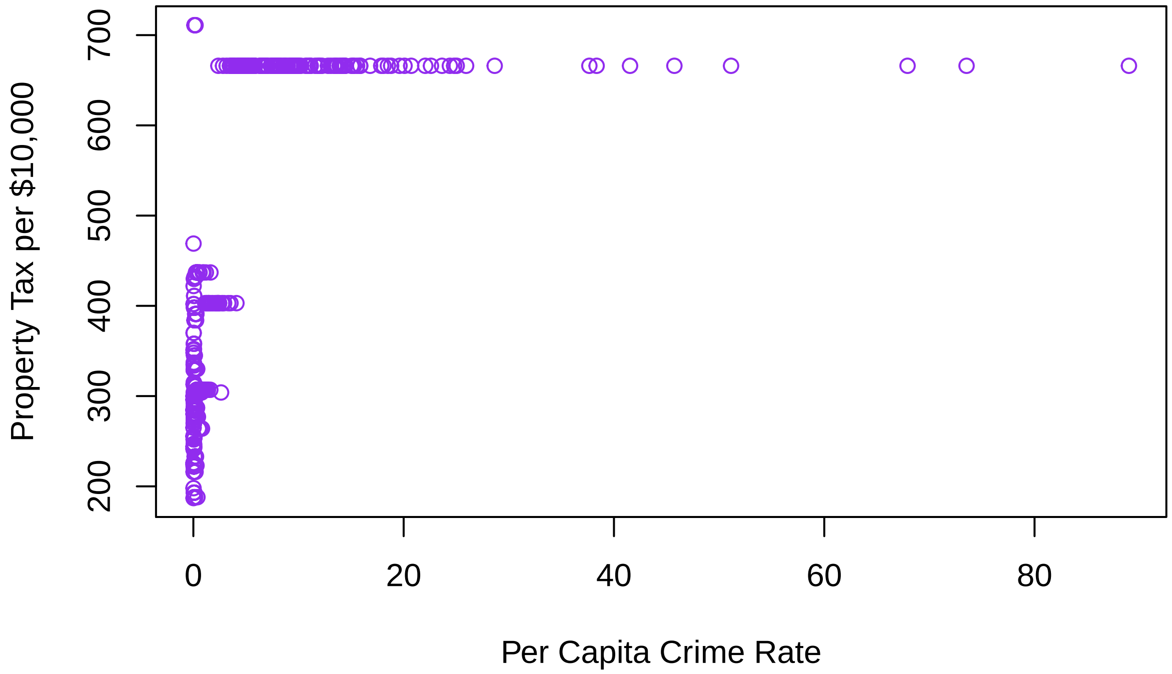
```
data(Boston)
library(ggplot2)
Boston1 <- Boston
plot(x = Boston1$dis, y = Boston1$tax, xlab = "Distance to Businesses", ylab = "Property Tax per $10,00
```



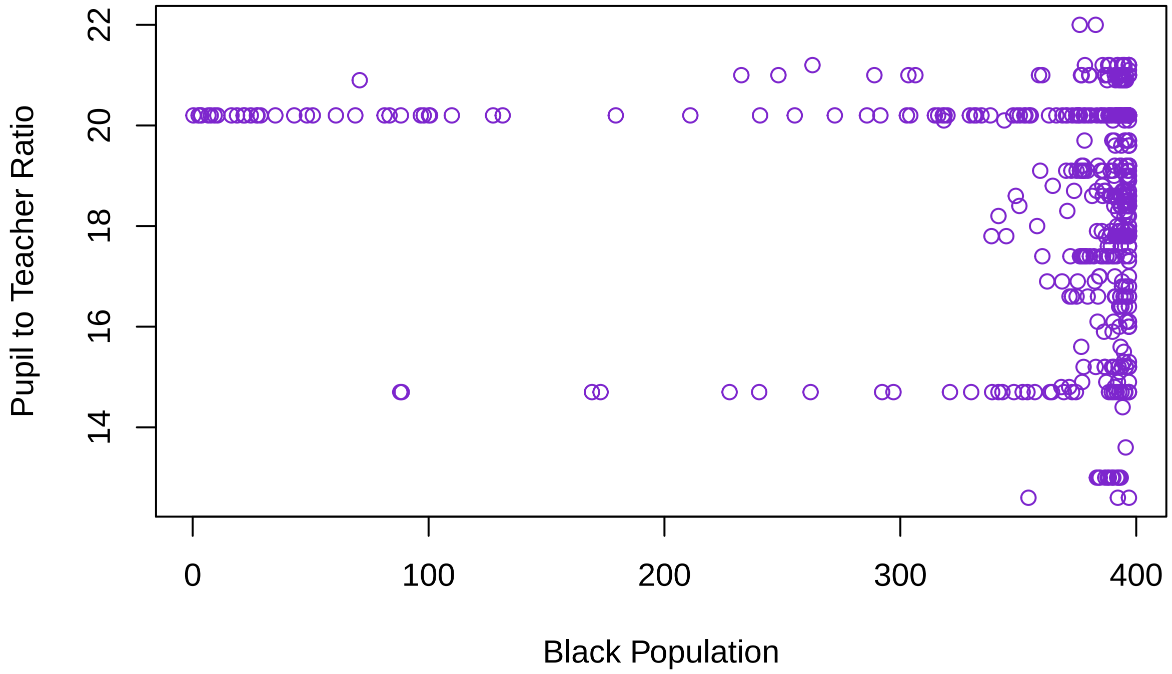
```
ggplot(Boston1, aes(x = dis, y = tax)) +
  geom_point(alpha = .6, color = "purple2")
```



```
plot(x = Boston1$crim, y = Boston1$tax, xlab = "Per Capita Crime Rate", ylab = "Property Tax per $10,00
```



```r
plot(x = Boston1$black, y = Boston1$ptratio, xlab = "Black Population", ylab = "Pupil to Teacher Ratio"
```



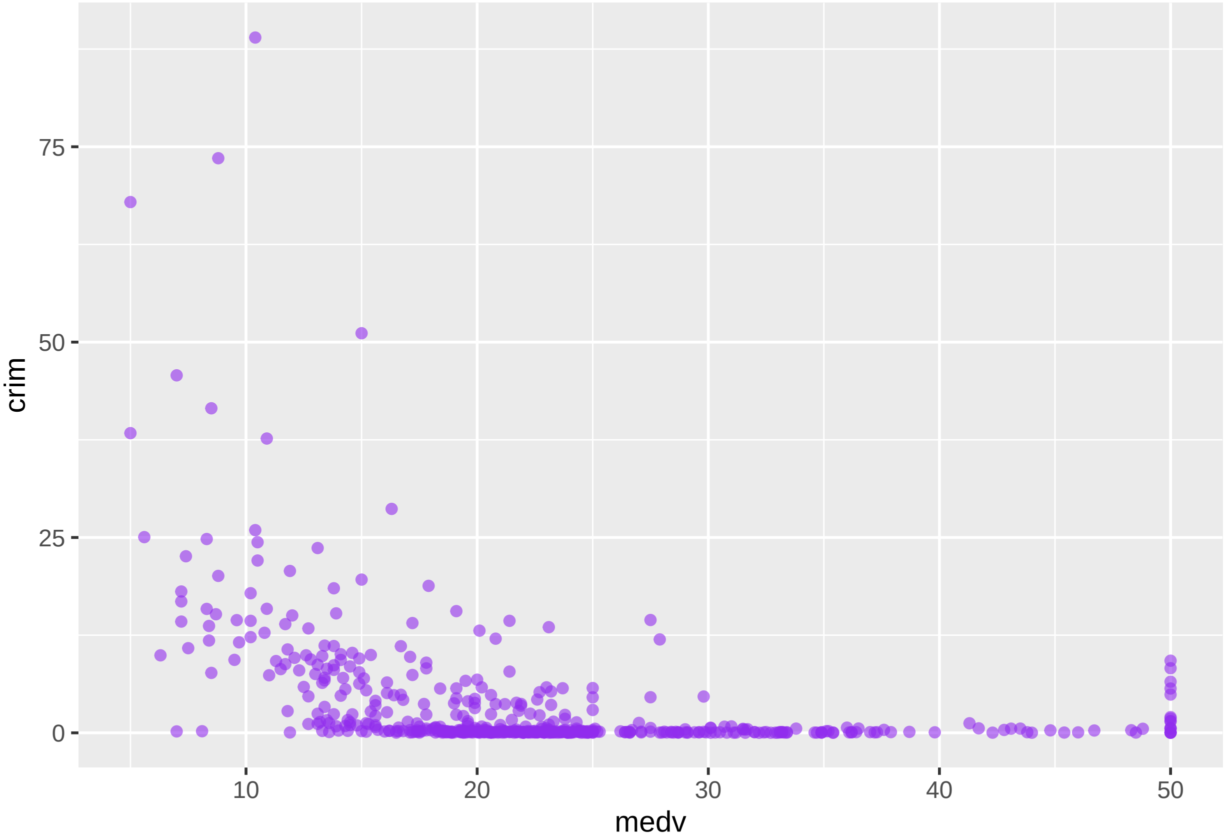
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
```

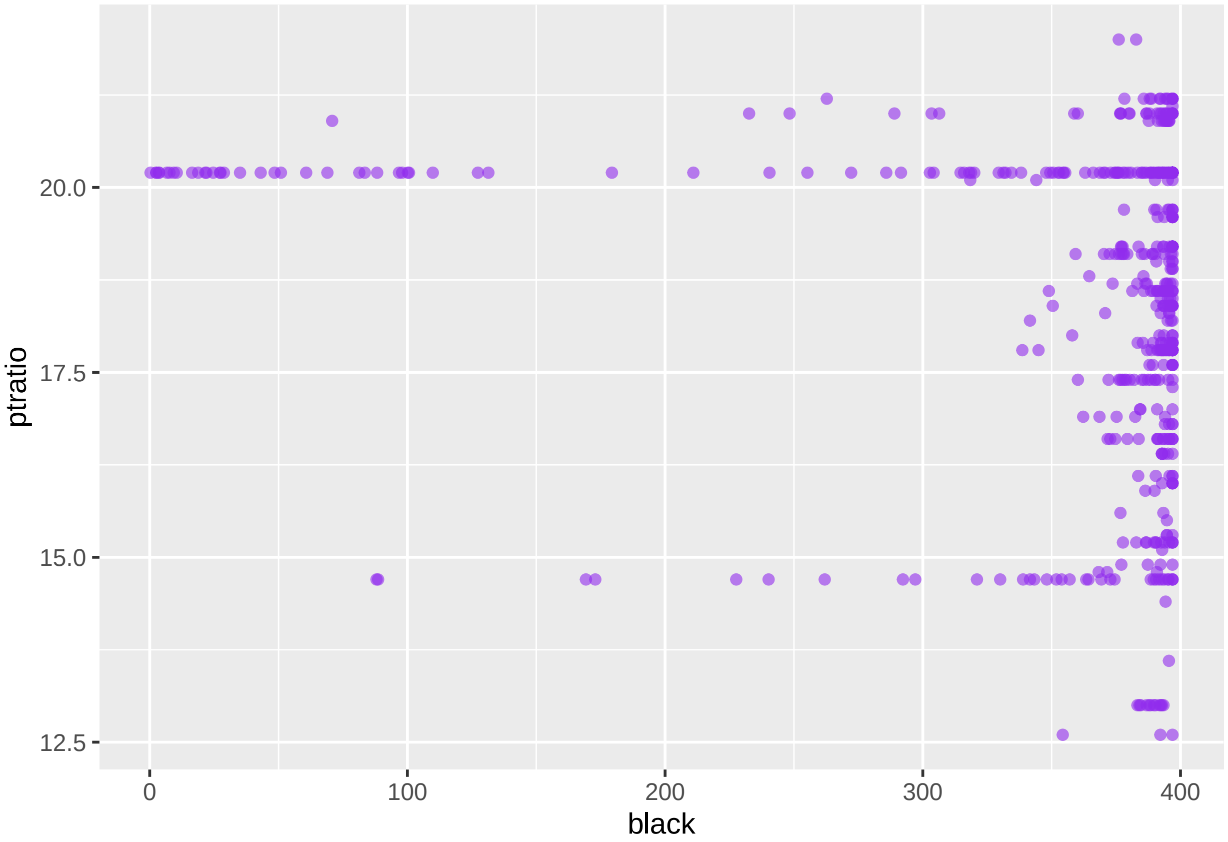
```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

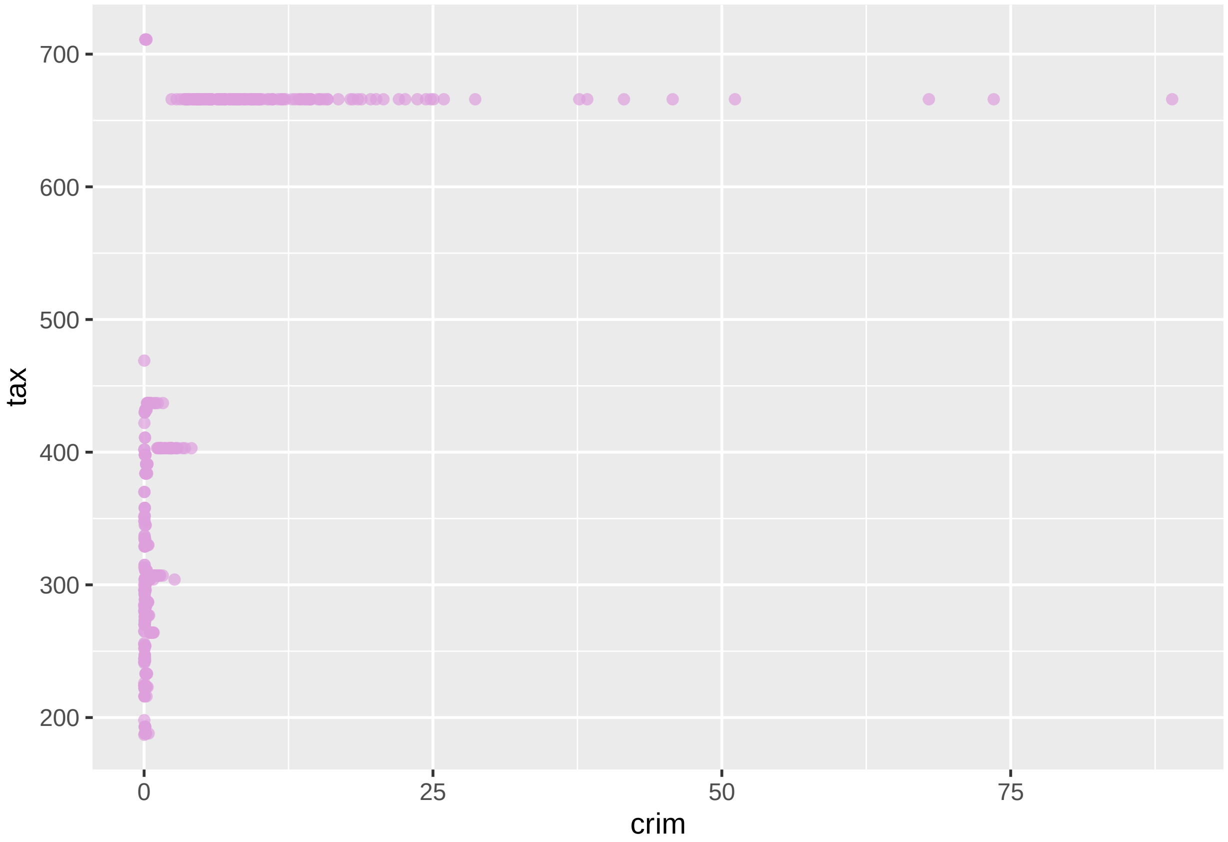
```
Boston2 <- Boston1 %>%
  mutate(blackprop = (((black/1000)^(.5))+.63))

library(ggplot2)
ggplot(data = Boston1, mapping = aes(x = medv, y = crim)) +
  geom_point(alpha = .6, col = "purple2")
```



```
ggplot(Boston1, aes(x = black, y = ptratio)) +
  geom_point(alpha = .6, col = "purple2")
```



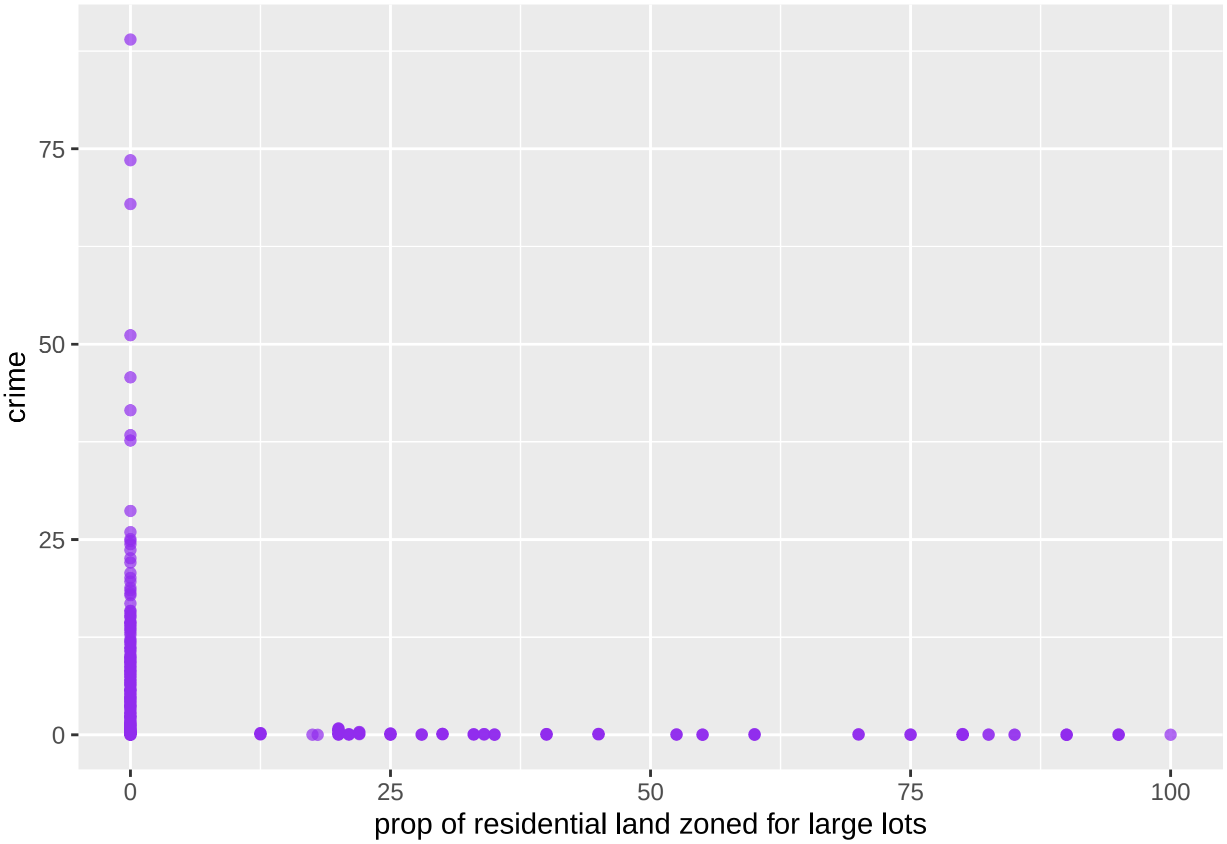
```r
ggplot(Boston1, aes(x = crim, y = tax)) +
  geom_point(alpha = .7, col = "plum")
```



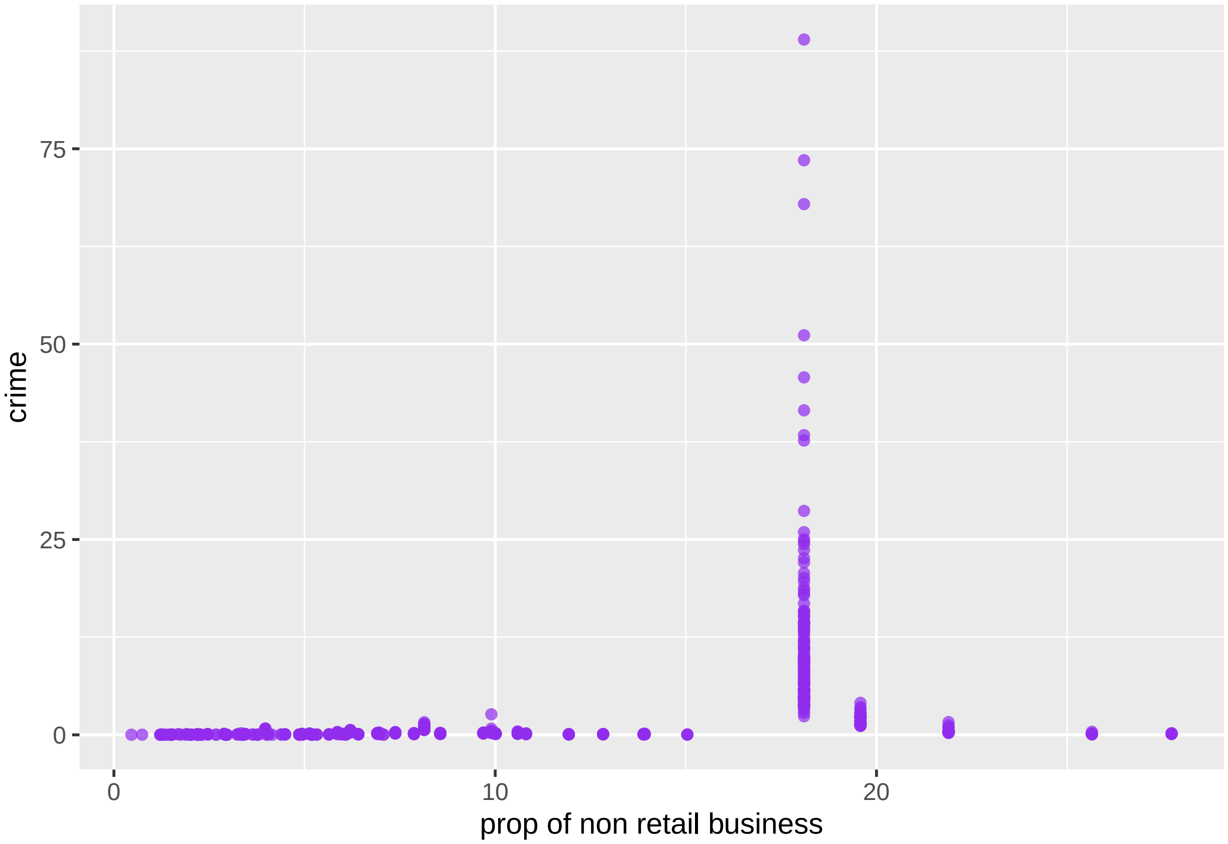
Some things I found through the making of these graphs: From the first ggplot, it seems that as the median value of owner occupied homes (medv) goes up, we see that the crime rate decreases until the very end where the values reach around 50k. Finally, in the last ggplot, we compare property tax to crime rates. It seems in the majority of towns, the property tax is below 5,000,000, and for these towns, the crime rate is relatively low. However, once we reach the 6,500,000 mark, the crime rate increases. The towns stay at a little bit above $6,500,000, but different towns have different crime rates around here, and they have a pretty wide range. Note: I had a graph or two that had the variable "black", but I don't feel comfortable trying to extract a meaning from these graphs because I do not understand how they derived "black" from the proportion of black people by town. I tried to reverse engineer the proportion of black people from the equation they gave "1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.", but my proportions ended being over 1 in my Boston2 Data. I can say that when the variable black started hiting the ~320 mark, the pupil to teacher ratio started to branch out from the 20 that it was stuck at for a while.

Exercise 3

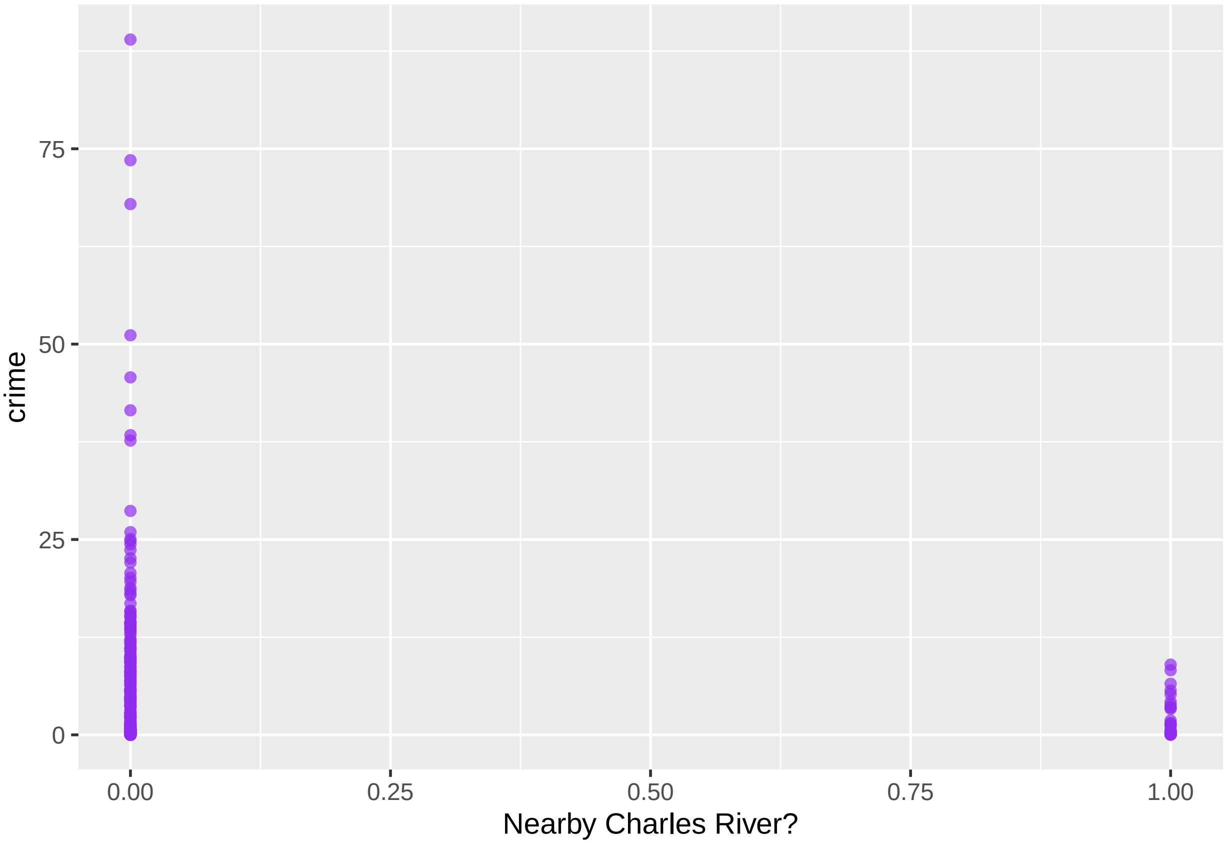
```
ggplot(data = Boston1, aes(x = zn, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("prop of residential land zoned for large lots") +
  ylab ("crime")
```



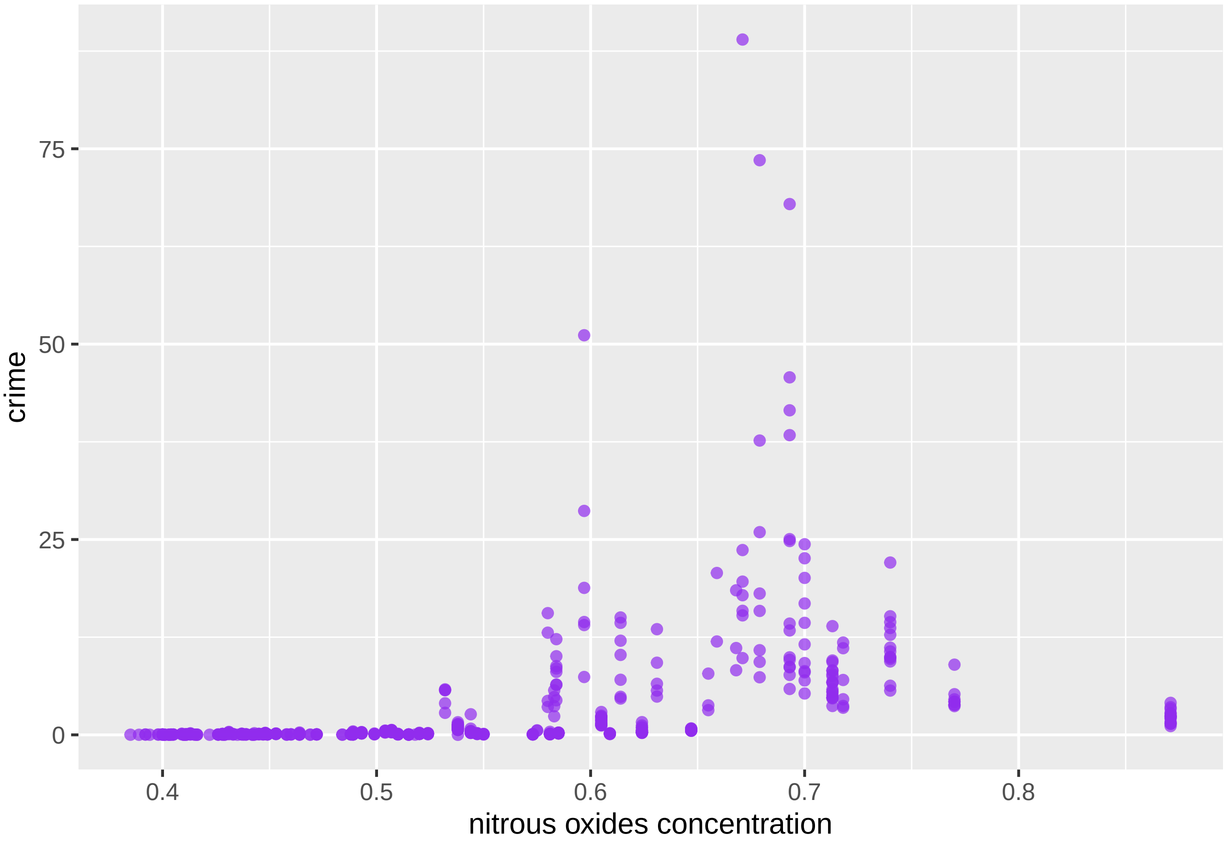
```
ggplot(data = Boston1, aes(x = indus, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("prop of non retail business") +
  ylab ("crime")
```



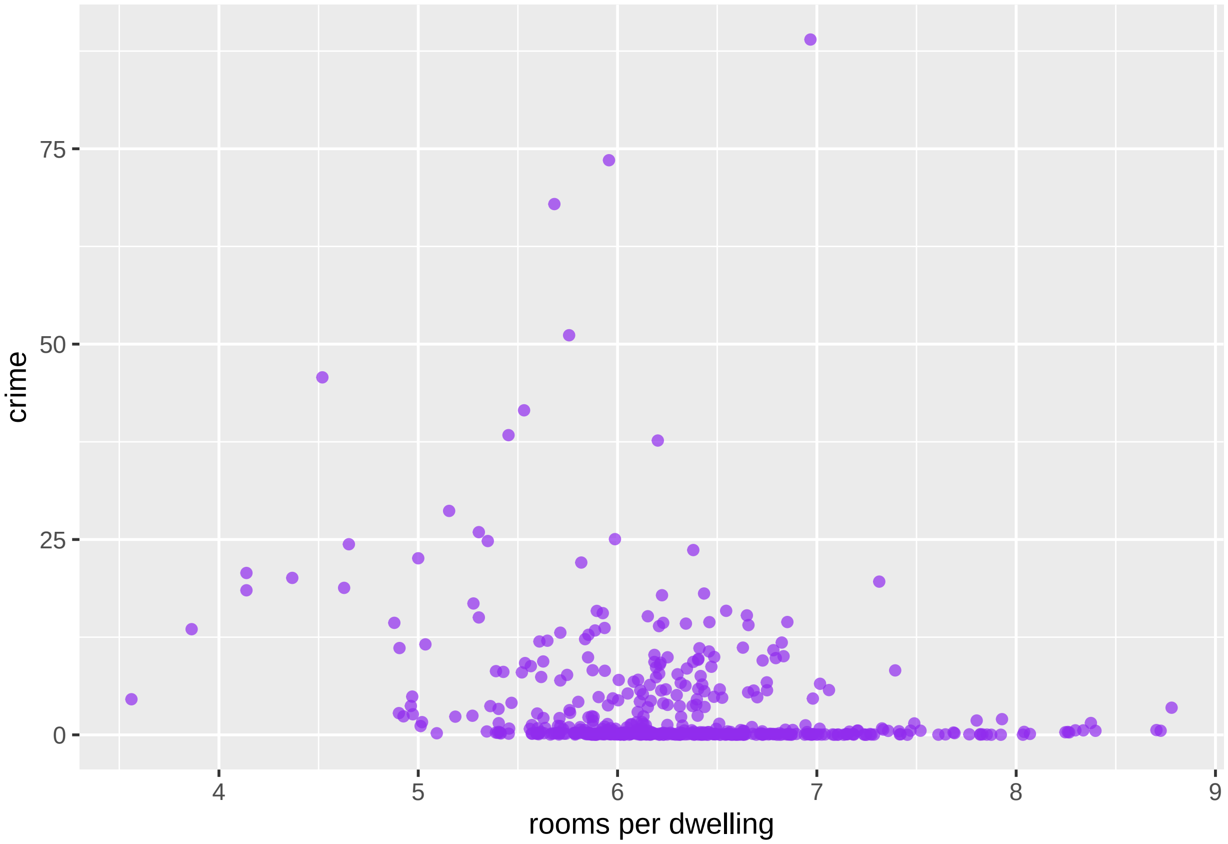
```
ggplot(data = Boston1, aes(x = chas, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("Nearby Charles River?") +
  ylab ("crime")
```



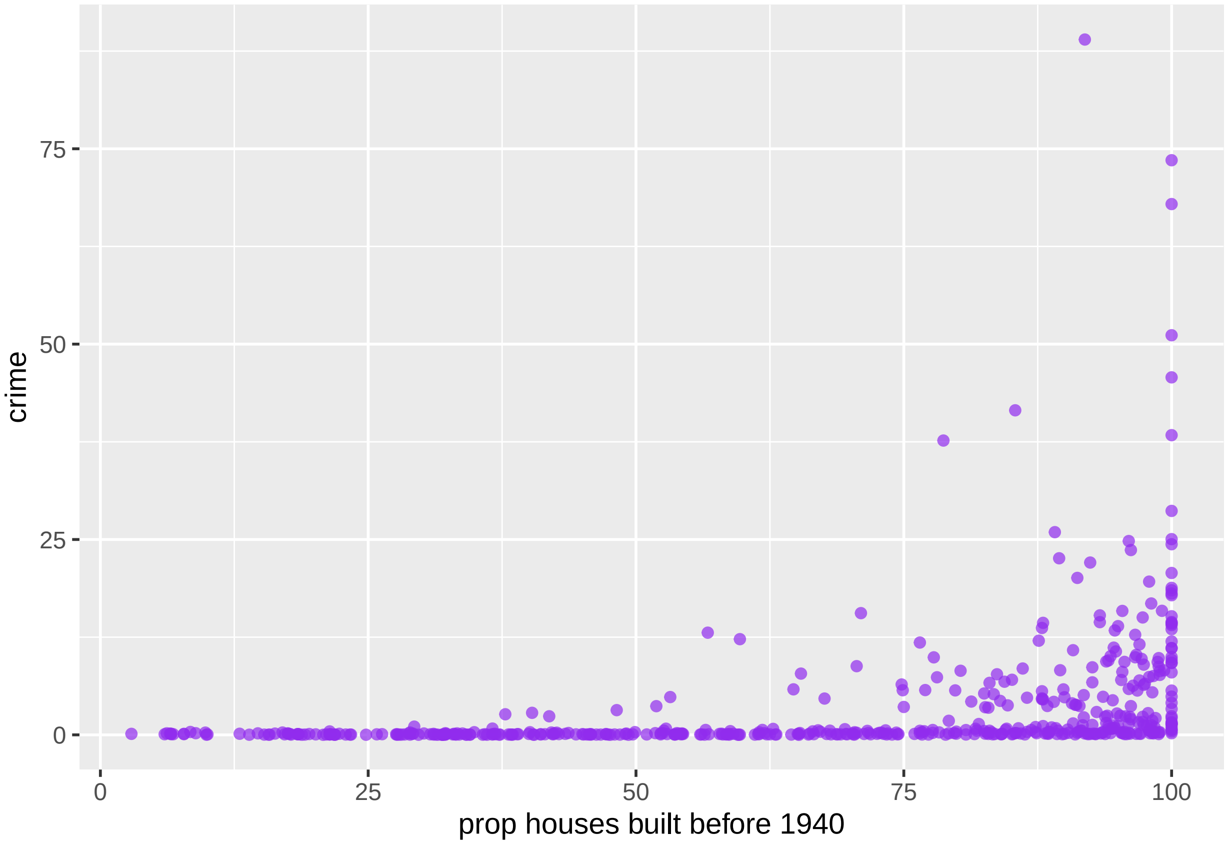
```
ggplot(data = Boston1, aes(x = nox, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("nitrous oxides concentration") +
  ylab ("crime")
```



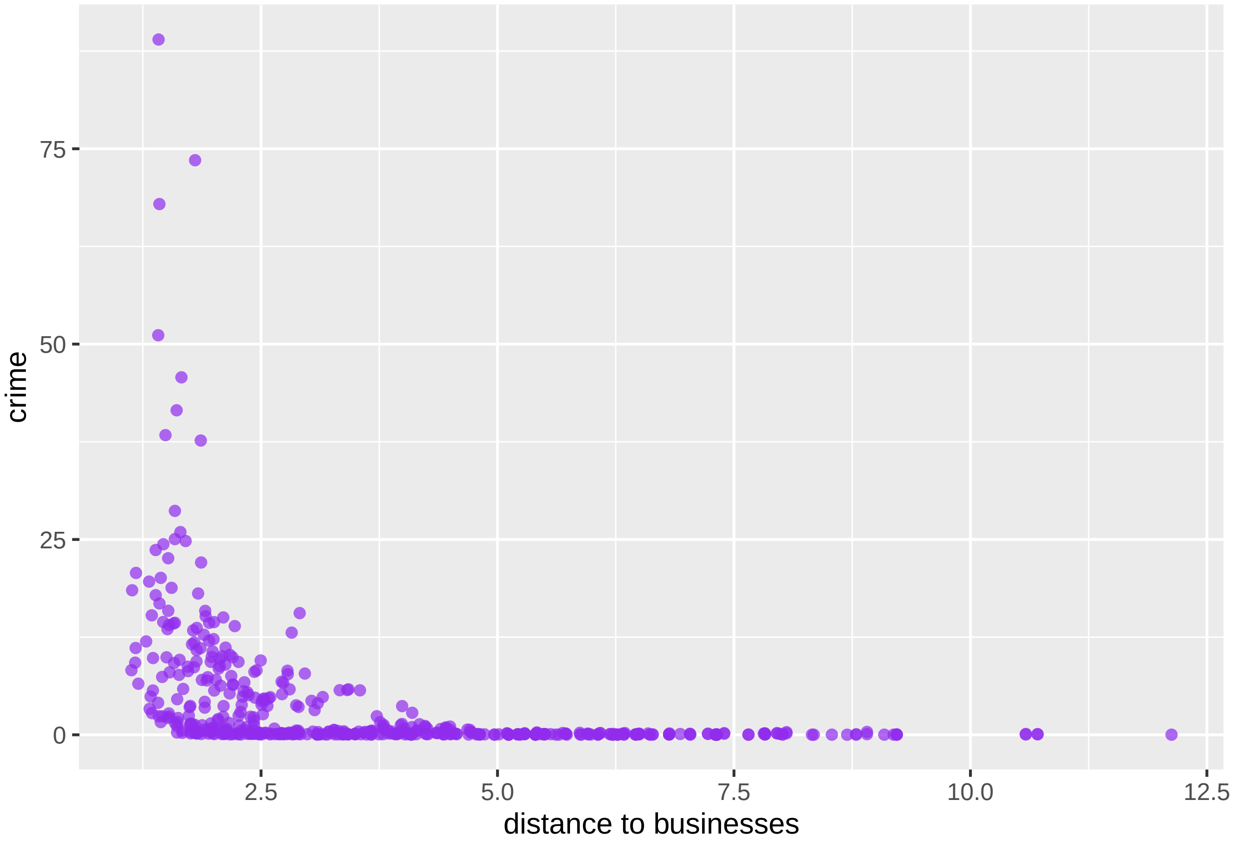
```
ggplot(data = Boston1, aes(x = rm, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("rooms per dwelling") +
  ylab ("crime")
```



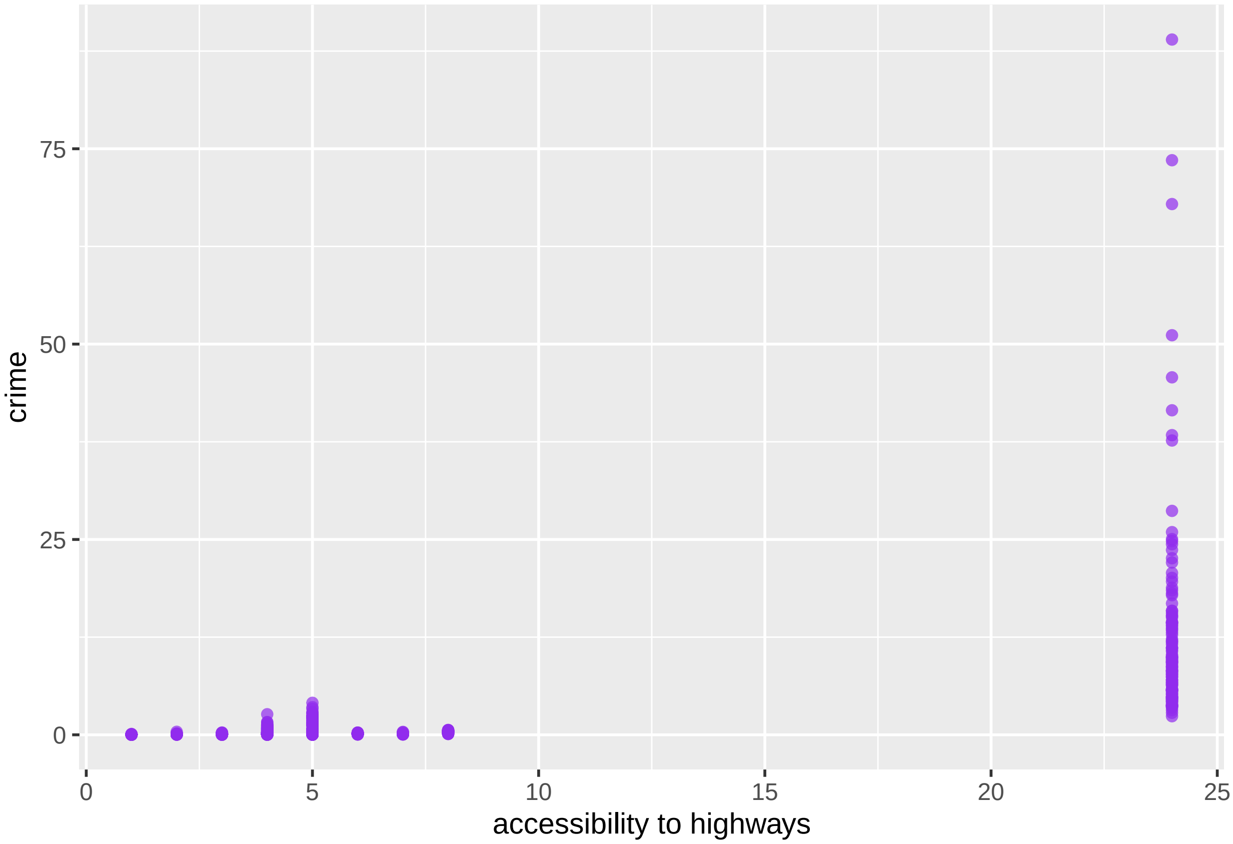
```
ggplot(data = Boston1, aes(x = age , y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("prop houses built before 1940") +
  ylab ("crime")
```



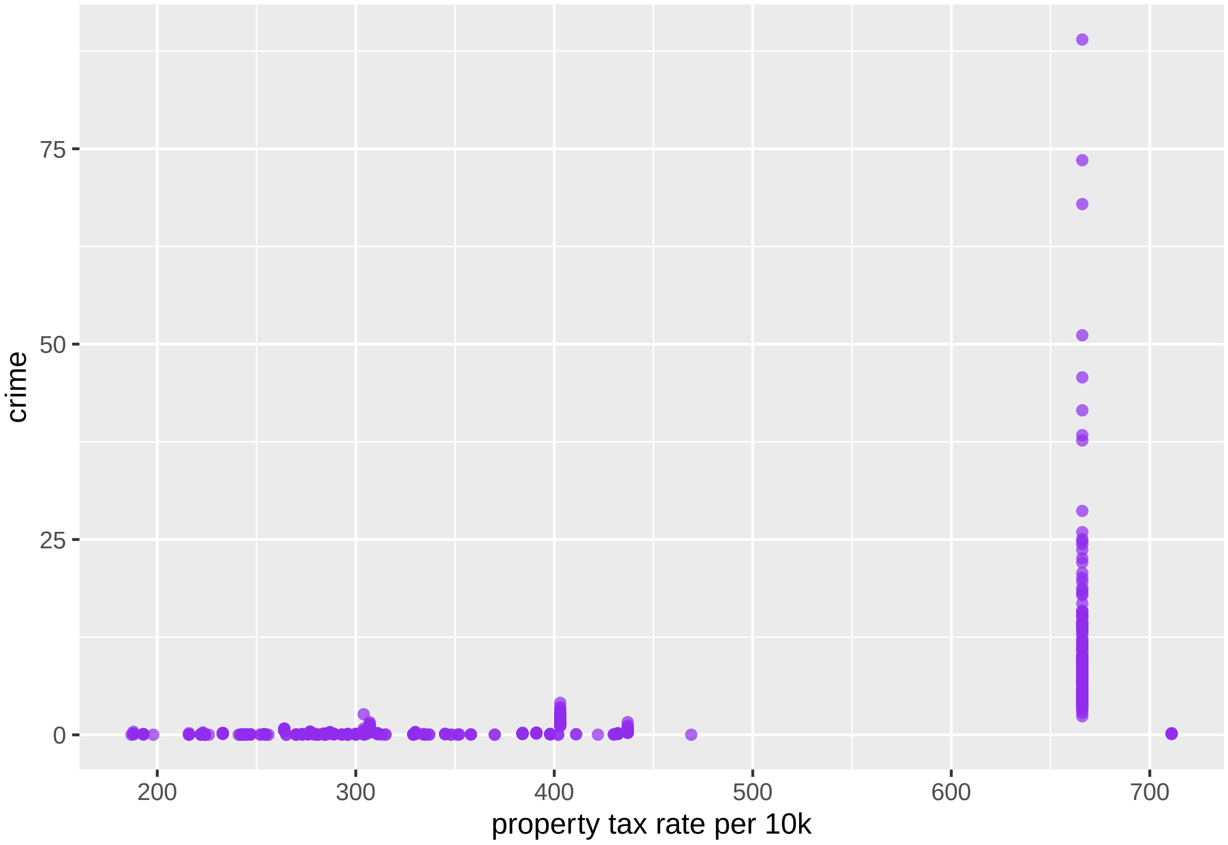
```
ggplot(data = Boston1, aes(x = dis, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("distance to businesses") +
  ylab ("crime")
```



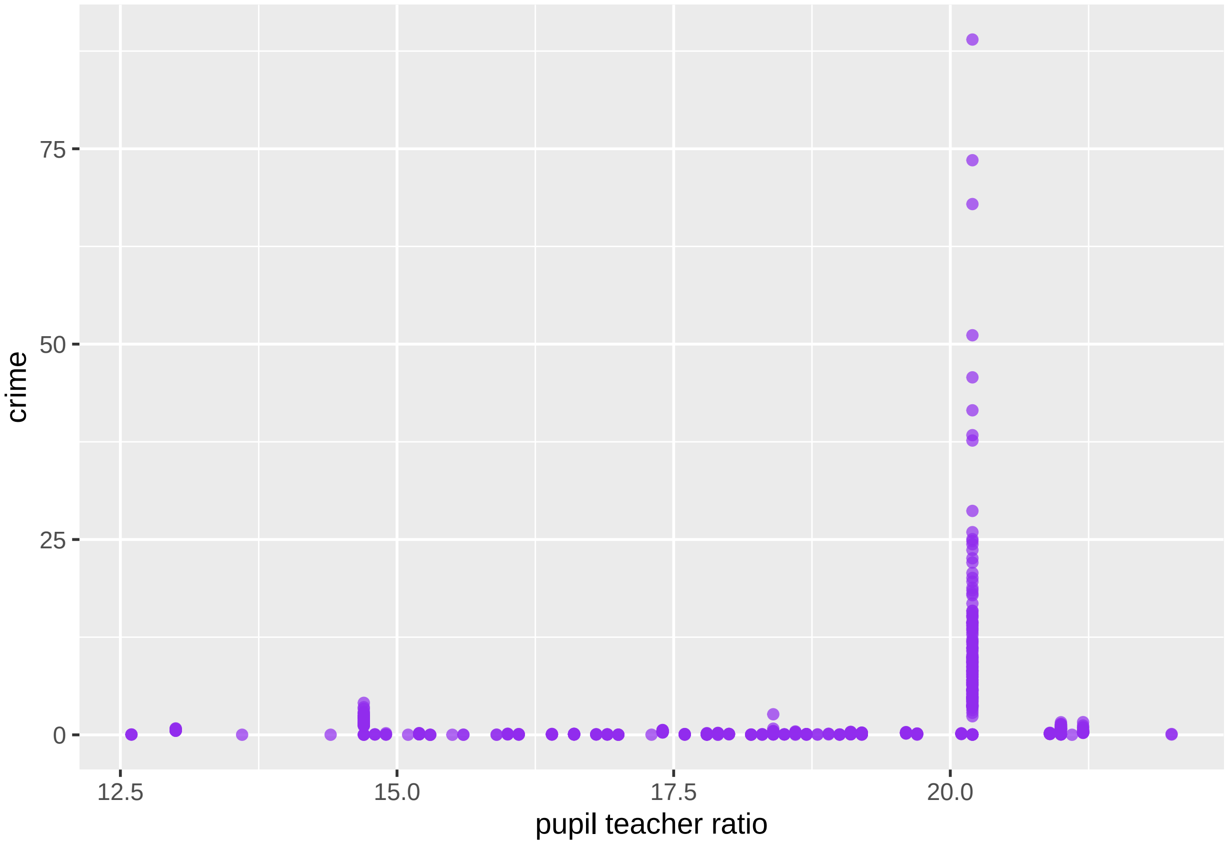
```
ggplot(data = Boston1, aes(x = rad, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("accessibility to highways") +
  ylab ("crime")
```



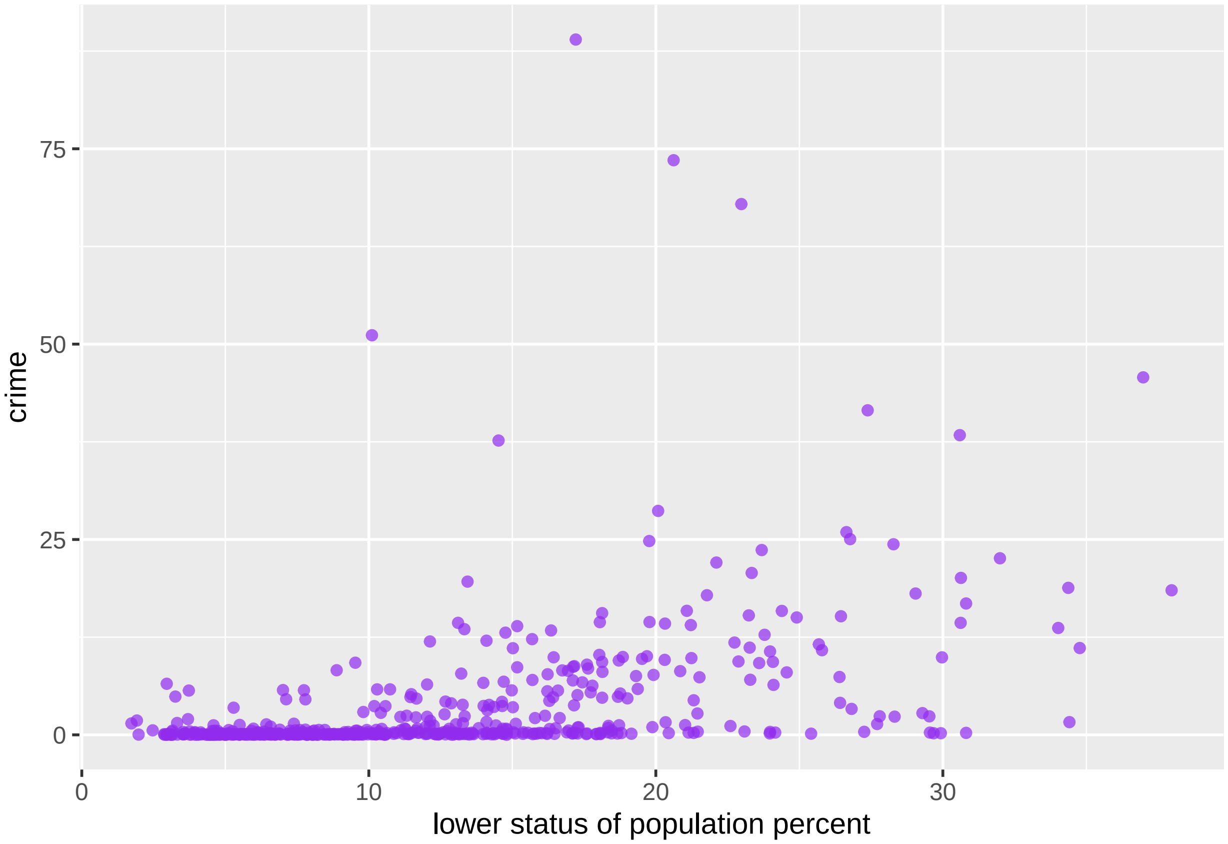
```
ggplot(data = Boston1, aes(x = tax, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("property tax rate per 10k") +
  ylab ("crime")
```



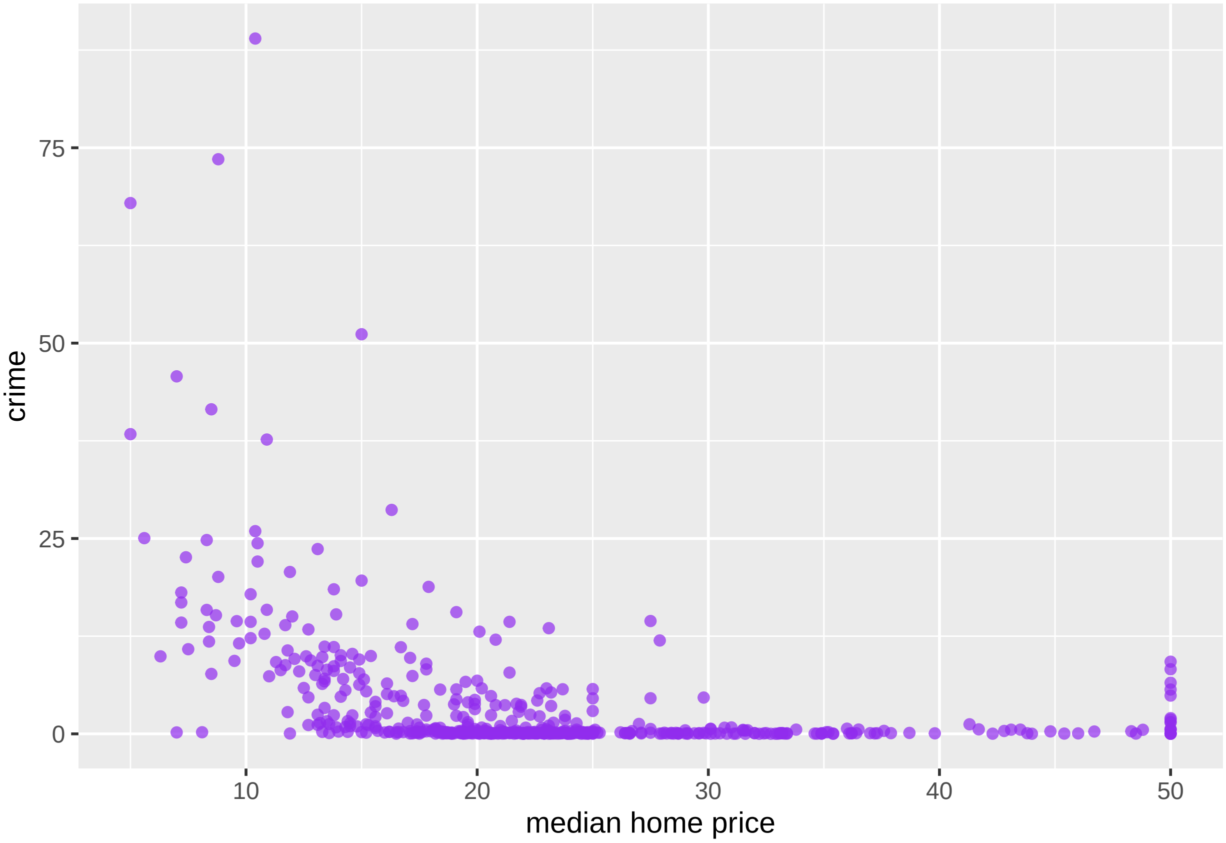
```
ggplot(data = Boston1, aes(x = ptratio, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("pupil teacher ratio") +
  ylab ("crime")
```



```
ggplot(data = Boston1, aes(x = lstat, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("lower status of population percent") +
  ylab ("crime")
```



```
ggplot(data = Boston1, aes(x = medv, y = crim)) +
  geom_point(color = "purple2", alpha = .7) +
  xlab ("median home price") +
  ylab ("crime")
```



I found that some predictors have a relationship with crime rate. The first one I noticed was nitrous oxides concentration. I noticed that as the concentration increased, the crime rate increased as well. Next, I noticed that when the average rooms per dwelling decreased, crime rate decreased. I also saw that the towns where the houses were more recent had lower crime rates than the towns with a high proportion of houses built before 1940. I also saw distance to businesses being important in regards to crime rate: the closest towns to boston's employment centers had higher crime rates than towns further away. The lower status graph also demonstrated that in towns where more people were considered "lower class", crime rate increased. In my last graph, I saw that higher median home prices are correlated with lower crime rates.

Exercise 4

```
range(Boston1$crim)
```

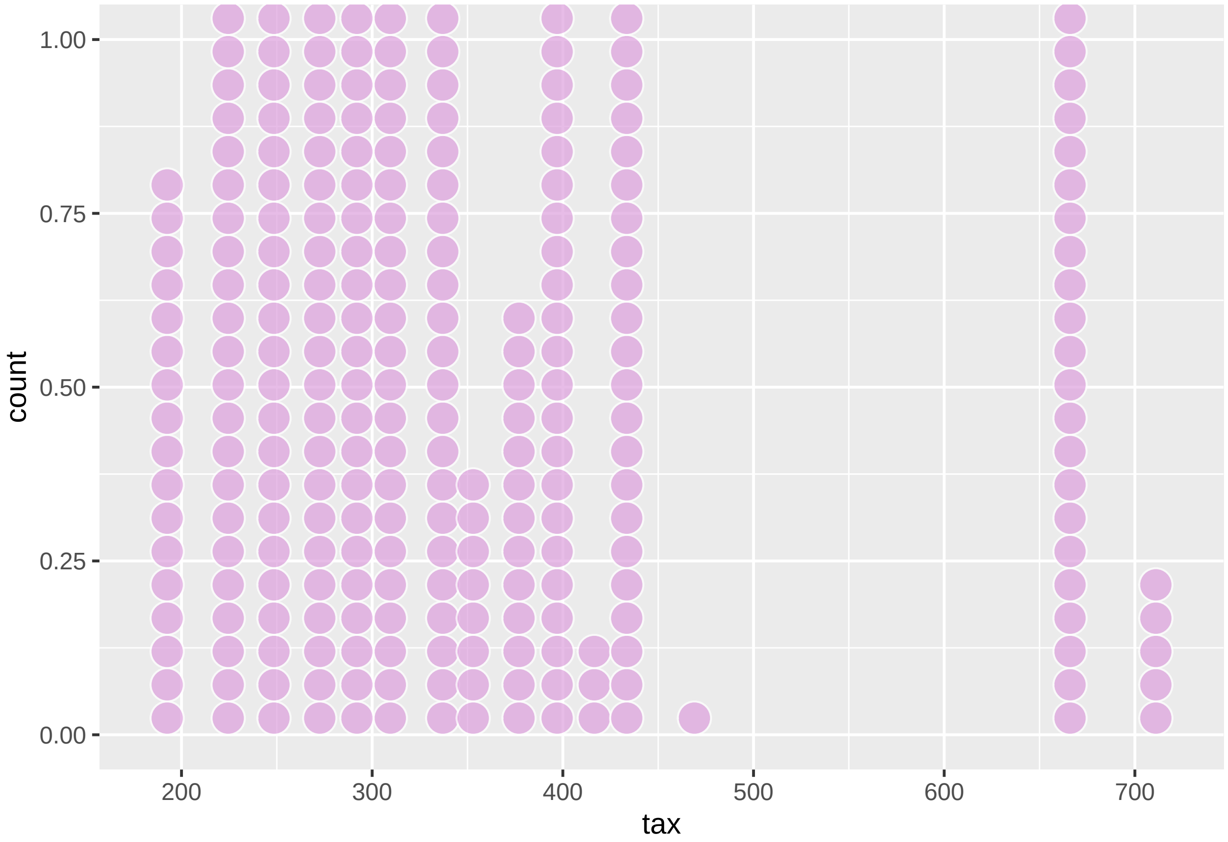
```
## [1]  0.00632 88.97620
```

```
#arrange(Boston1, desc(crim))


range(Boston1$tax)
```

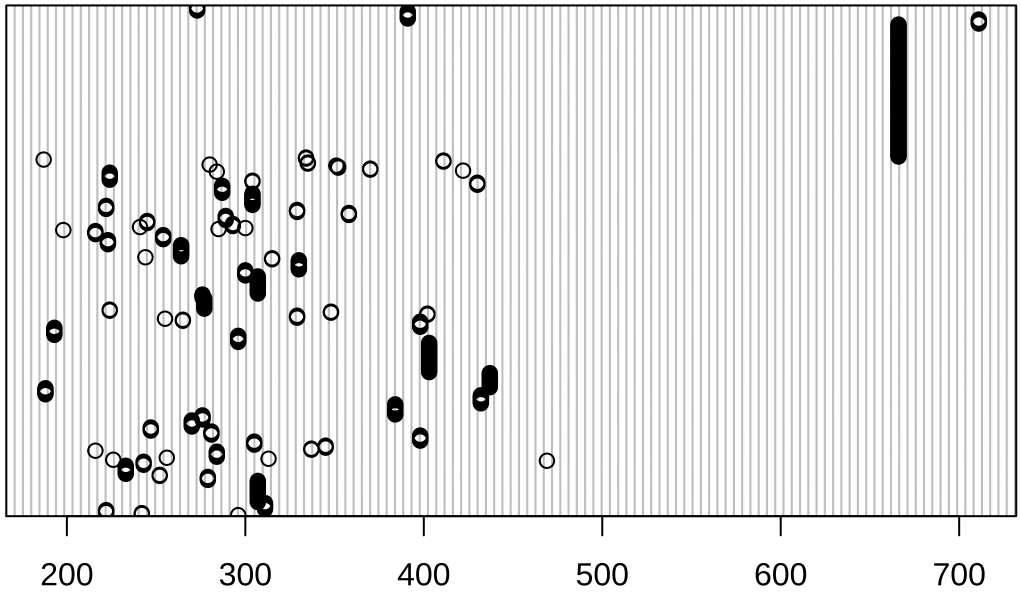
```
## [1] 187 711
```

```
#arrange(Boston1, desc(tax))
ggplot(Boston1, aes(x = tax)) +
  geom_dotplot(alpha = .7, col = "white", fill = "plum")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
dotchart(Boston1$tax)
```

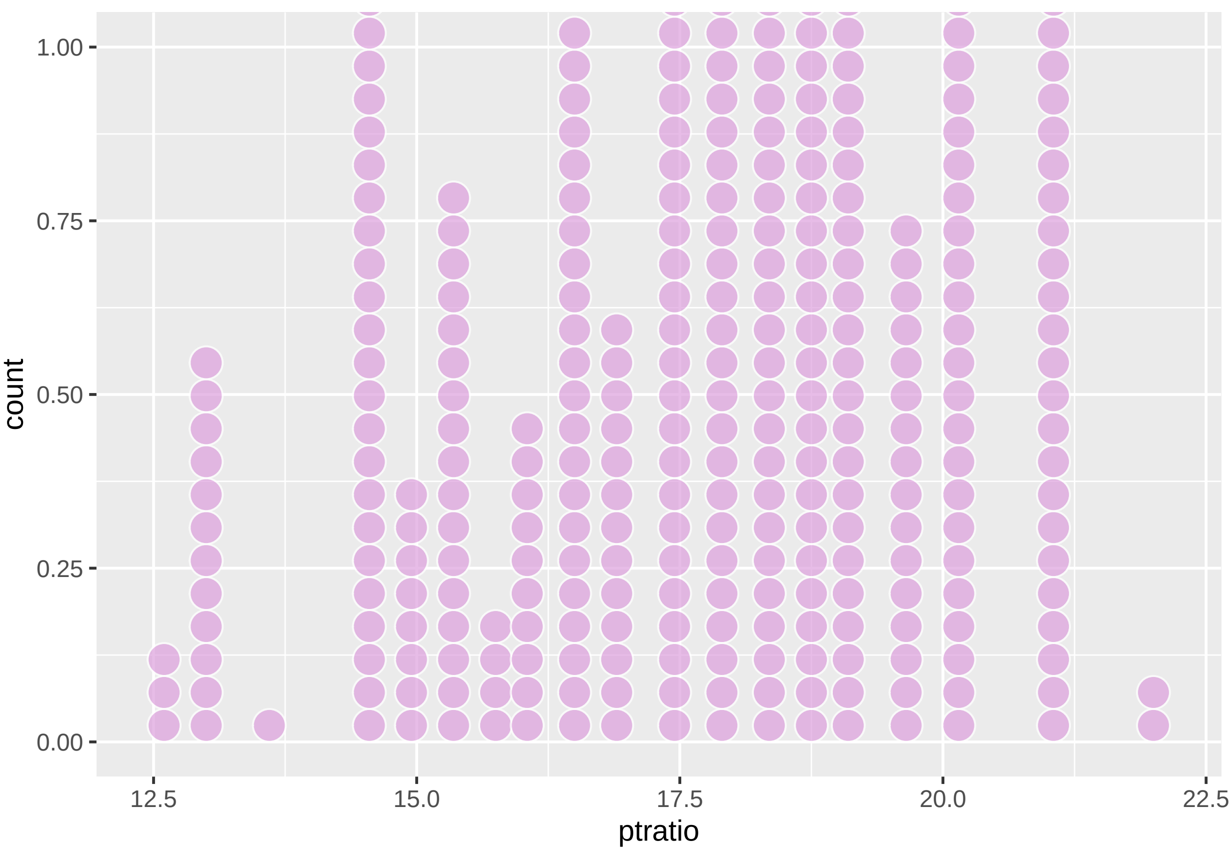


```r
range(Boston1$ptratio)
```

```
## [1] 12.6 22.0
```

```r
#arrange(Boston1, desc(ptratio))
ggplot(Boston1, aes(x = ptratio)) +
  geom_dotplot(alpha = .7, col = "white", fill = "plum")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



There are some suburbs with high crime rate, with some suburbs being as high as 88.98, 73.53, but with many super low also (.00632). I know sometimes per capita is multiplied by 100k so that its not really small, but for there to be towns with .0063 percent, I feel like this is not the case, so this seems to be a pretty large range, to go from almost 0 to 80.

There are five suburbs with tax rates of 711, but there are a bunch of suburbs where the tax is 666. However past that, there is a big gap between the next suburbs, and then theres a lot of variation once you hit the 460~ mark. I would say this range is pretty large. The property tax from 711-187 is a pretty different level of standard of living.

For pupil teacher ratio, I wouldnt say there are any suburbs with especially high ratios, the highest are at 22, and then it pretty gradually drops down to 12.5. I don't really think that range for this is that high either.. The difference between a 20 kid classroom and a 12 isnt that bad.

Exercise 5

```r
sum(Boston1$chas)
```

```
## [1] 35
```

There are 35 suburbs next to the Charles River.

Exercise 6

```r
median(Boston1$ptratio)
```

```
## [1] 19.05
```

The median pupil teacher ratio is 19.05 students to a teacher

Exercise 7

```
#medvcor <- function(pred) {
 # correlation <- cor(Boston1$medv, Boston1$pred)
  #return(correlation)
#}

#cor(Boston1$medv, Boston1$crim)
#medvcor(crim)
#not sure why doesnt work

cor(Boston1$medv, Boston1$crim)
```

```
## [1] -0.3883046
```

```
cor(Boston1$medv, Boston1$zn)
```

```
## [1] 0.3604453
```

```
cor(Boston1$medv, Boston1$indus)
```

```
## [1] -0.4837252
```

```
cor(Boston1$medv, Boston1$chas)
```

```
## [1] 0.1752602
```

```
cor(Boston1$medv, Boston1$nox)
```

```
## [1] -0.4273208
```

```
cor(Boston1$medv, Boston1$rad)
```

```
## [1] -0.3816262
```

```
cor(Boston1$medv, Boston1$age)
```

```
## [1] -0.3769546
```

```
cor(Boston1$medv, Boston1$dis)
```

```
## [1] 0.2499287
```

```
cor(Boston1$medv, Boston1$rad)
```

```
## [1] -0.3816262
```

```
cor(Boston1$medv, Boston1$tax)
```

```
## [1] -0.4685359
```

```
cor(Boston1$medv, Boston1$ptratio)
```

```
## [1] -0.5077867
```

```
cor(Boston1$medv, Boston1$black)
```

```
## [1] 0.3334608
```

```
cor(Boston1$medv, Boston1$lstat)
```

```
## [1] -0.7376627
```

The model's response would be the median value of the home of a suburb. Our inputs would be the values of the suburb's different variables: crim, zn, indus, nox, rad, age, dis, tax, ptratio, black, and lstat. I wouldnt include whether they are next to the Charles River or if the Distance from employment centers because the strength of the relationship between median homevalue and these variables are not as powerful when compared to the others.