Homework 1

Manav Mandhani, mm58926

1. (1 point) Consider two binary vectors u and v. Suppose the total number of ones in both the binary vectors together is n; and that dot product of the two vectors is d. What is the Jaccard similarity of u and v?

d/n

2. (1 point) Let ei denote the vector of length n which has 1 at index i and 0 elsewhere. What is the cosine similarity between the vectors ei and ej when (i) i = j? (ii) i = j?

If i = j, cosine similarity is 1/1 = 1 If i != j, cosine similarity is 0/1 = 0

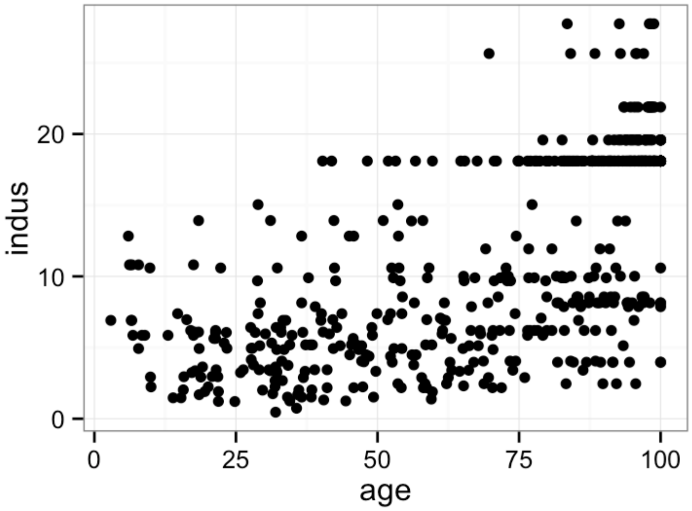
a. How many rows and columns are there in this data set? What do the rows and columns represent?

The data has 506 rows and 14 columns.

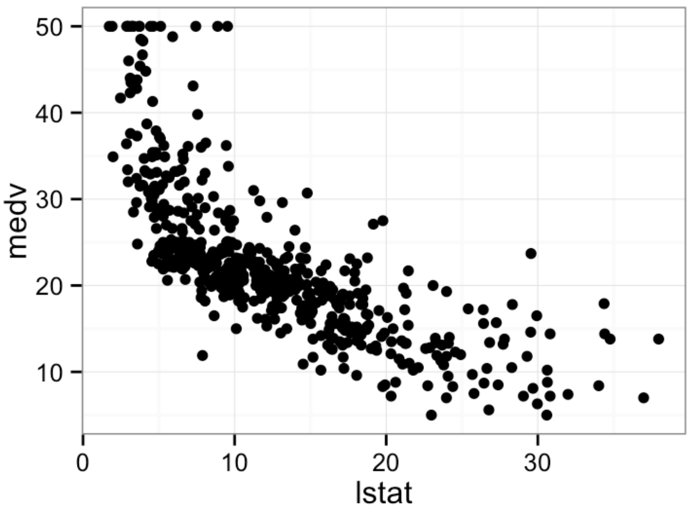
Each row represents one object unit representing one suburb and the columns represent different housing values for each suburb

b. Make some pairwise scatterplots of the attributes (columns) in this data set. Describe your findings.

```
ggplot(Boston, aes(x = age, y = indus)) + geom_point()
```



```
ggplot(Boston, aes(x = lstat, y = medv)) + geom_point()
```



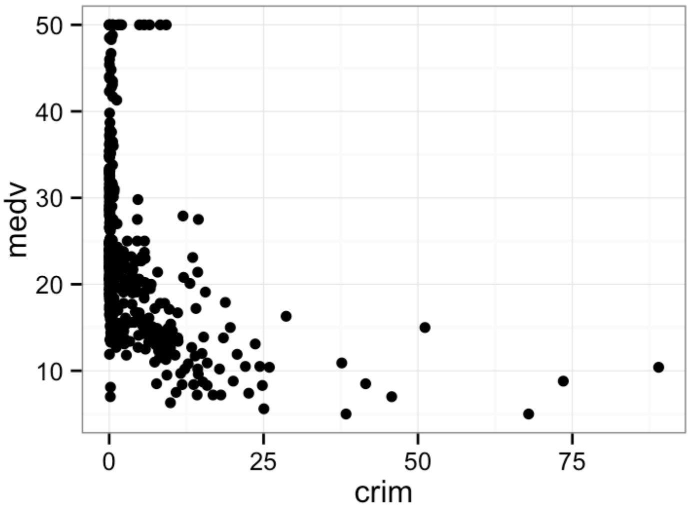
For both these plots there is some sort of a dependency between the attributes.

For the first plot, as the age increases, the proportion of non-retail business acres remains constant until a certain age but increaeses after a certain age.

For the second plot, the median value of owner-occupied homes gradually decreases as the percent of population in lower-status increases.

c. From the scatterplots, can you identify any attribute that is associated with per-capita crime rate?

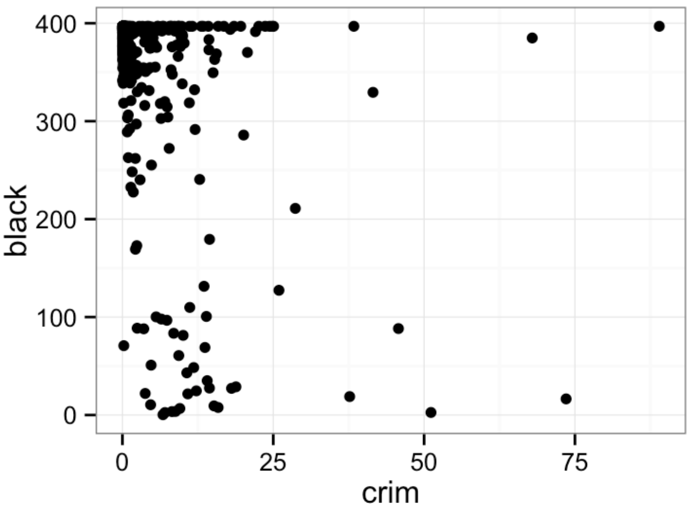
```
ggplot(Boston, aes(x = crim, y = medv)) + geom_point()
```



There isn’t a strong association with any such attribute but crime is seen to increases as the median value of owner-occupied homes decreases

d. Do any of the suburbs of Boston appear to have particularly high crime rates? Do any have particularly high pupil-teacher ratios?

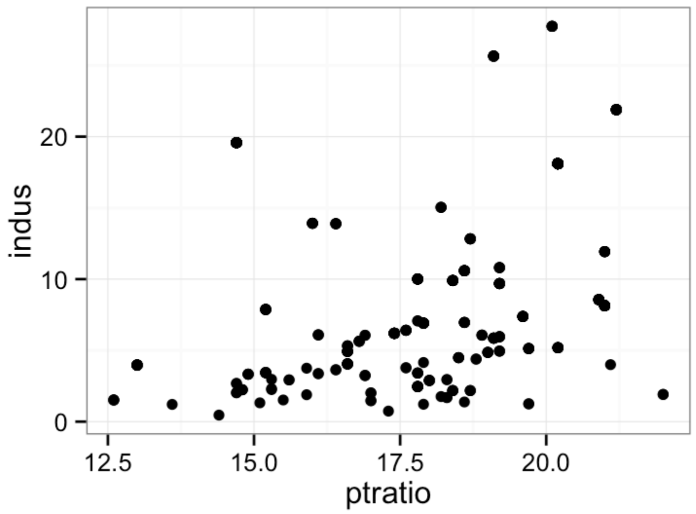
```
ggplot(Boston, aes(x = crim, y = black)) + geom_point()
```



```
Boston %>% arrange(desc(crim)) -> boston.crime
slice(boston.crime, 1:10)
```

```
##       crim zn indus chas   nox    rm   age    dis rad tax ptratio  black
## 1  88.9762  0  18.1    0 0.671 6.968  91.9 1.4165  24 666    20.2 396.90
## 2  73.5341  0  18.1    0 0.679 5.957 100.0 1.8026  24 666    20.2  16.45
## 3  67.9208  0  18.1    0 0.693 5.683 100.0 1.4254  24 666    20.2 384.97
## 4  51.1358  0  18.1    0 0.597 5.757 100.0 1.4130  24 666    20.2   2.60
## 5  45.7461  0  18.1    0 0.693 4.519 100.0 1.6582  24 666    20.2  88.27
## 6  41.5292  0  18.1    0 0.693 5.531  85.4 1.6074  24 666    20.2 329.46
## 7  38.3518  0  18.1    0 0.693 5.453 100.0 1.4896  24 666    20.2 396.90
## 8  37.6619  0  18.1    0 0.679 6.202  78.7 1.8629  24 666    20.2  18.82
## 9  28.6558  0  18.1    0 0.597 5.155 100.0 1.5894  24 666    20.2 210.97
## 10 25.9406  0  18.1    0 0.679 5.304  89.1 1.6475  24 666    20.2 127.36
##    lstat medv
## 1  17.21 10.4
## 2  20.62  8.8
## 3  22.98  5.0
## 4  10.11 15.0
## 5  36.98  7.0
## 6  27.38  8.5
## 7  30.59  5.0
## 8  14.52 10.9
## 9  20.08 16.3
## 10 26.64 10.4
```

```
ggplot(Boston, aes(x = ptratio, y = indus)) + geom_point()
```



```
Boston %>% arrange(desc(ptratio)) -> boston.ptratio
slice(boston.ptratio, 1:10)
```

```
##       crim zn indus chas   nox    rm  age     dis rad tax ptratio  black
## 1  0.04301 80  1.91    0 0.413 5.663 21.9 10.5857   4 334    22.0 382.80
## 2  0.10659 80  1.91    0 0.413 5.936 19.5 10.5857   4 334    22.0 376.04
## 3  0.25915  0 21.89    0 0.624 5.693 96.0  1.7883   4 437    21.2 392.11
## 4  0.32543  0 21.89    0 0.624 6.431 98.8  1.8125   4 437    21.2 396.90
## 5  0.88125  0 21.89    0 0.624 5.637 94.7  1.9799   4 437    21.2 396.90
## 6  0.34006  0 21.89    0 0.624 6.458 98.9  2.1185   4 437    21.2 395.04
## 7  1.19294  0 21.89    0 0.624 6.326 97.7  2.2710   4 437    21.2 396.90
## 8  0.59005  0 21.89    0 0.624 6.372 97.9  2.3274   4 437    21.2 385.76
## 9  0.32982  0 21.89    0 0.624 5.822 95.4  2.4699   4 437    21.2 388.69
## 10 0.97617  0 21.89    0 0.624 5.757 98.4  2.3460   4 437    21.2 262.76
##    lstat medv
## 1   8.05 18.2
## 2   5.57 20.6
## 3  17.19 16.2
## 4  15.39 18.0
## 5  18.34 14.3
## 6  12.60 19.2
## 7  12.26 19.6
## 8  11.12 23.0
## 9  15.03 18.4
## 10 17.31 15.6
```

Yes, there are certain neighborhoods that have rates as high as 88.97 and 73.53 which is much more than the average crime rate.

However, there seems to be an even spread on pupil teacher ratio across suburbs

e. How many suburbs in this data set bound the Charles river?

```
Boston %>% filter(chas == 1) %>% nrow()
```

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

35 suburbs

f. What is the median pupil-teacher ratio among the towns in this data set?

```
Boston %>% summarize(median.ptr = median(ptratio))
```

```
##   median.ptr
## 1      19.05
```

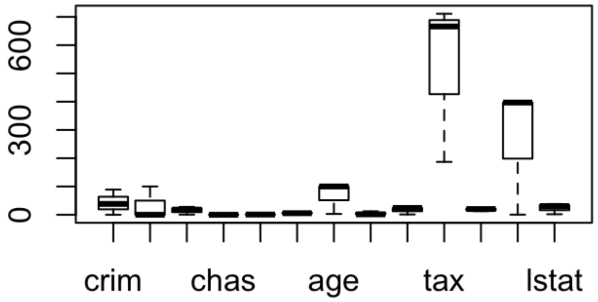
Median pupil-teacher ratio is 19.05

g. In this dataset, which suburb of Boston has the lowest median value of owner-occupied homes? What are the values of the other attributes for this suburb, and how do those values compare to the overall range for those attributes?

```
Boston %>% arrange(medv) %>% slice(1) -> boston.lowest.median
boston.lowest.median
```

```
##      crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
## 1 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
##   medv
## 1    5
```

```
boston.lowest.median <- boston.lowest.median[, -14]
crim <- c(range(Boston$crim)[1], range(Boston$crim)[2])
zn <- c(range(Boston$zn)[1], range(Boston$zn)[2])
indus <- c(range(Boston$indus)[1], range(Boston$indus)[2])
chas <- c(range(Boston$chas)[1], range(Boston$chas)[2])
nox <- c(range(Boston$nox)[1], range(Boston$nox)[2])
rm <- c(range(Boston$rm)[1], range(Boston$rm)[2])
age <- c(range(Boston$age)[1], range(Boston$age)[2])
dis <- c(range(Boston$dis)[1], range(Boston$dis)[2])
rad <- c(range(Boston$rad)[1], range(Boston$rad)[2])
tax <- c(range(Boston$tax)[1], range(Boston$tax)[2])
ptratio <- c(range(Boston$ptratio)[1], range(Boston$ptratio)[2])
black <- c(range(Boston$black)[1], range(Boston$black)[2])
lstat <- c(range(Boston$lstat)[1], range(Boston$lstat)[2])
frames <- data.frame(crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, l
stat)
framed <- rbind(boston.lowest.median, frames)
boxplot(framed, range = 1.5)
```



The boxplot shows the given value in the range of all other attributes. For attributes such as age, percentage of black population, it seems to have a higher value in the range. For others, such as proportion of residential land zoned for lots over 25,000 sq.ft, weighted mean of distances to five Boston employment centres, it seems to have been in the lower end of the range.

h. In this dataset, how many of the suburbs average more than eight rooms per dwelling? What can you say about these suburbs?

```
Boston %>% filter(rm >= 8) %>% nrow()
```

```
## [1] 13
```

13 suburbs average more than eight rooms per dwelling