PSY T880: HW 1

Tinashe Michael Tapera

24 January 2017

This exercise involves the Boston Dataset. To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

The data set is contained in the object Boston.

```
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
```

You can read about the data set using the call `?Boston`.

```
?Boston
```

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

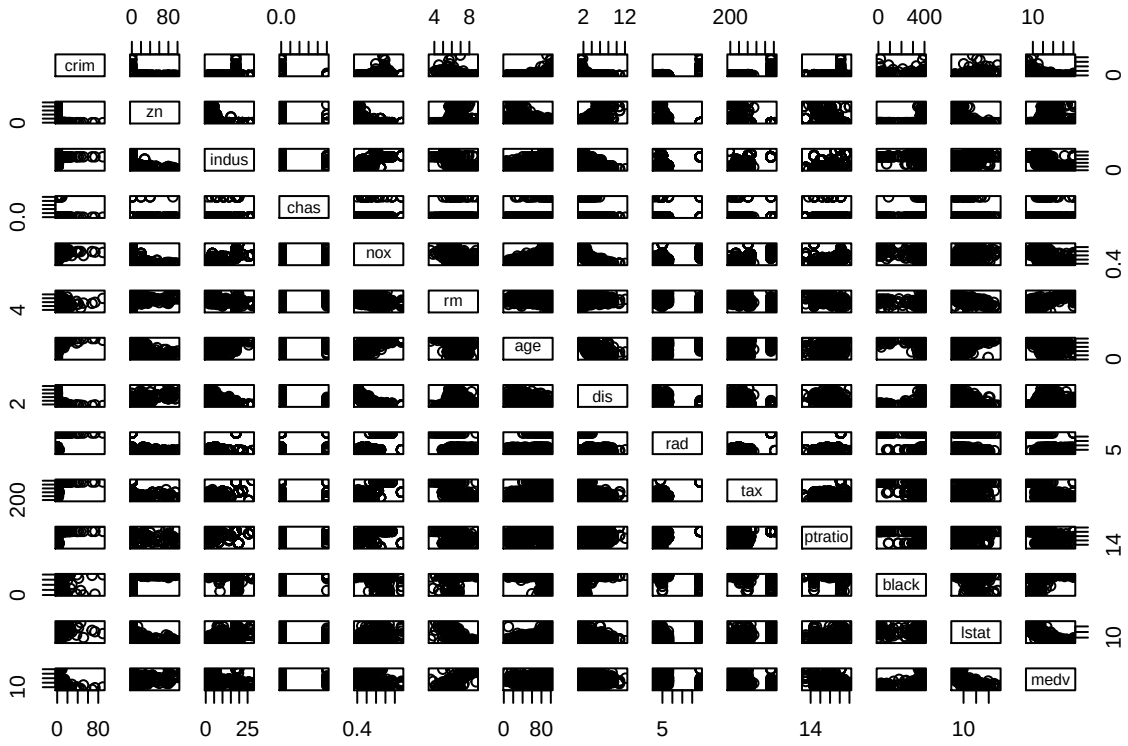
```
dim(Boston)
```

```
## [1] 506  14
```

The data set contains 506 rows and 14 columns, where rows represent observations and columns represent variables.

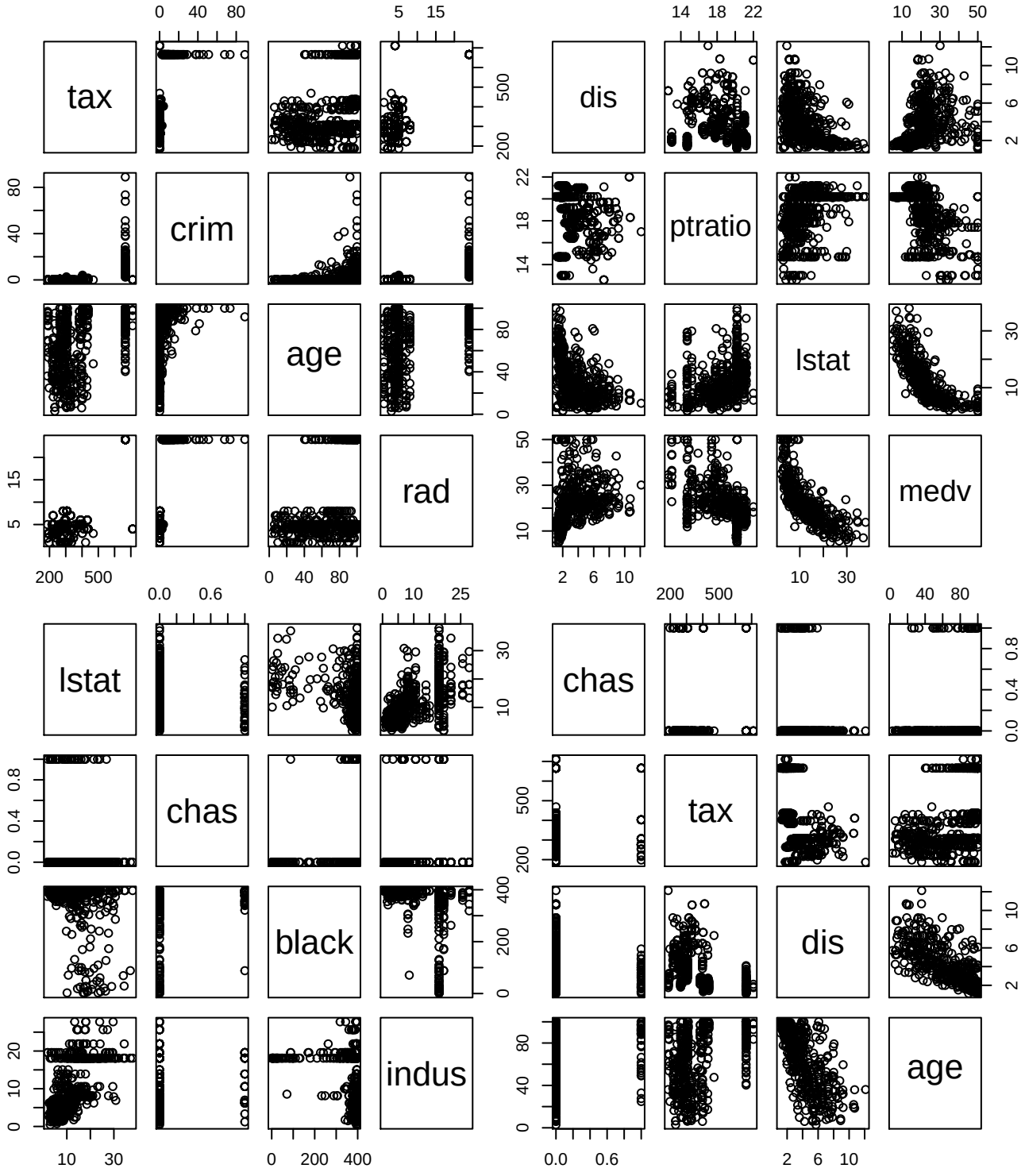
Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

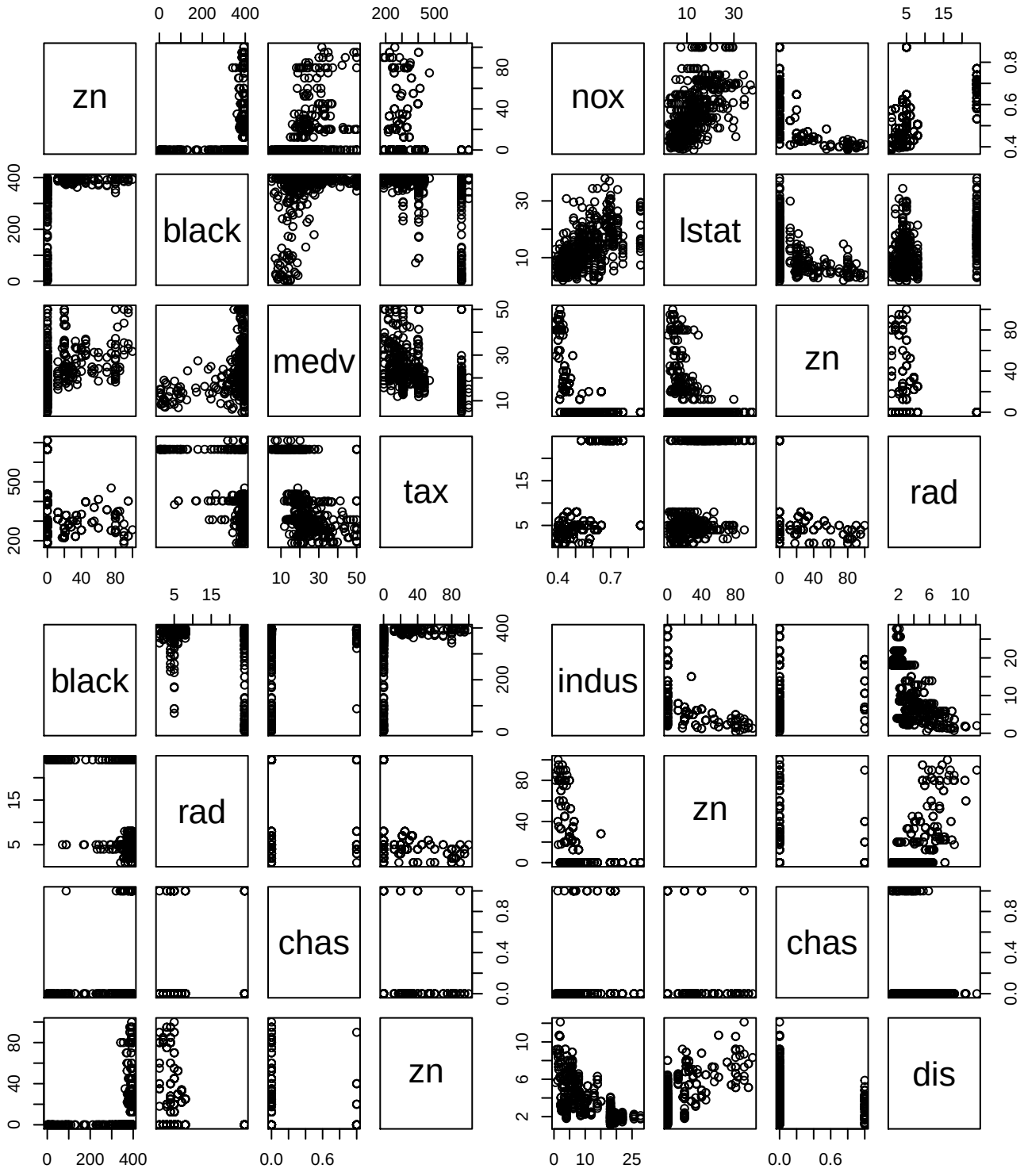
```
pairs(Boston)
```



This scatter plot is too dense to be able to gain much from. However, we can see that "chas" is a dichotomous variable (Charles River) and that "rad" is probably not normally distributed (Radial highways). We can subset this plot to see more clearly.

```
# take a random sample of 4 columns each time and
# use this to fill in the plot with a few randomly
# selected variables; 8 times is enough
for(i in 1:8){pairs(Boston[,sample(1:14,4)])}
```


tax
crim
age
rad
dis
ptratio
lstat
medv
lstat
chas
tax
black
dis
indus
age



Some more observations:

- The variable Tax seems to have some inflated value around 680, as well as Rad around 24, and Indus at around 19. zn (Zone) is probably zero inflated.
- Age may be positively correlated with lstat.
- Age may be positively correlated with nox
- zn may be correlated with rm
- dis may be correlated with zn

- crim and dis may be non-linearly related. There is a very clear law, perhaps exponential decay law, between the two.
- medv and rm may be linearly related
- medv may be correlated with lstat, although it may be non-linear, perhaps logarithmic decay.

  What is the predictor in the data set most correlated with per capita crime rate? Explain the relationship.

We could visually inspect each scatter plot, but it would be more efficient to just run a loop of correlation tests, where we run a correlation between each variable and the target variable, and note down the coefficient and its p-value.

```
correlations = data.frame(variable = NULL, coef = NULL, significant = NULL)
for(i in 2:dim(Boston)[2]){
  test = cor.test(Boston$crim,Boston[,i])
  variable = names(Boston)[i]
  coef = test$estimate
  significant = ifelse(test$p.value < 0.05, "YES", "NO")
  correlations = rbind(correlations,data.frame(variable,coef,significant, row.names = NULL))
}
kable(correlations, row.names = FALSE)
```

variable	coef	significant
zn	-0.2004692	YES
indus	0.4065834	YES
chas	-0.0558916	NO
nox	0.4209717	YES
rm	-0.2192467	YES
age	0.3527343	YES
dis	-0.3796701	YES
rad	0.6255051	YES
tax	0.5827643	YES
ptratio	0.2899456	YES
black	-0.3850639	YES
lstat	0.4556215	YES
medv	-0.3883046	YES

It looks like all of the variables except the chas variable are significantly correlated with crim. The chas variable is a categorical variable, so this makes sense. We can also order the observations to see which is most strongly correlated.

```
kable(correlations[order(-abs(correlations$coef)),], row.names = FALSE)
```

variable	coef	significant
rad	0.6255051	YES
tax	0.5827643	YES
lstat	0.4556215	YES
nox	0.4209717	YES
indus	0.4065834	YES
medv	-0.3883046	YES
black	-0.3850639	YES
dis	-0.3796701	YES
age	0.3527343	YES
ptratio	0.2899456	YES

variable	coef	significant
rm	-0.2192467	YES
zn	-0.2004692	YES
chas	-0.0558916	NO

From this table we can see that rad (accessibility to radial highways) is most strongly correlated with crim(per capita crime rate), followed by tax (full value property-tax rate per $10k) and lstat (% lower status of the population).

How many of the suburbs in this data set bound the Charles River?

```
table(Boston$chas)
```

```
##
##   0   1
## 471  35
```

The table function counts how many of each observation there is in the variable. In this case, 0 represents Not bounding, and 1 represents Bounds. There are 35 suburbs that bound the river.

What is the median pupil-teacher ratio among the towns of this data set?

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

The median pupil-teacher ratio is 19.05 (or just 19).

Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
head(Boston[order(Boston$medv),])
```

```
##         crim zn indus chas   nox    rm   age    dis rad tax ptratio  black
## 399 38.35180  0 18.10    0 0.693 5.453 100.0 1.4896  24 666    20.2 396.90
## 406 67.92080  0 18.10    0 0.693 5.683 100.0 1.4254  24 666    20.2 384.97
## 401 25.04610  0 18.10    0 0.693 5.987 100.0 1.5888  24 666    20.2 396.90
## 400  9.91655  0 18.10    0 0.693 5.852  77.8 1.5004  24 666    20.2 338.16
## 415 45.74610  0 18.10    0 0.693 4.519 100.0 1.6582  24 666    20.2  88.27
## 490  0.18337  0 27.74    0 0.609 5.414  98.3 1.7554   4 711    20.1 344.05
##     lstat medv
## 399 30.59  5.0
## 406 22.98  5.0
## 401 26.77  5.6
## 400 29.97  6.3
## 415 36.98  7.0
## 490 23.97  7.0
```

```
apply(Boston, 2, summary)
```

```
##             crim     zn indus    chas    nox    rm    age    dis    rad
## Min.     0.00632   0.00  0.46 0.00000 0.3850 3.561   2.90  1.130  1.000
## 1st Qu.  0.08204   0.00  5.19 0.00000 0.4490 5.886  45.02  2.100  4.000
## Median   0.25650   0.00  9.69 0.00000 0.5380 6.208  77.50  3.207  5.000
## Mean     3.61400  11.36 11.14 0.06917 0.5547 6.285  68.57  3.795  9.549
## 3rd Qu.  3.67700  12.50 18.10 0.00000 0.6240 6.624  94.07  5.188 24.000
```

```
## Max.    88.98000 100.00 27.74 1.00000 0.8710 8.780 100.00 12.130 24.000
##           tax ptratio  black lstat  medv
## Min.    187.0   12.60   0.32  1.73  5.00
## 1st Qu. 279.0   17.40 375.40  6.95 17.02
## Median  330.0   19.05 391.40 11.36 21.20
## Mean    408.2   18.46 356.70 12.65 22.53
## 3rd Qu. 666.0   20.20 396.20 16.96 25.00
## Max.    711.0   22.00 396.90 37.97 50.00
```

The suburbs 399 and 406 both have the minimum median value at 5.0; We'll focus on suburb 399.

When compared to the values produced by a `summary()` call, we see that 399 has the following attributes:

- Greater than 3rd quartile criminality rate
- 0 land zoned lots over 25k sq. ft. (minimum for variable)
- 18.1% of businesses are non-retail, at exactly the 3rd quartile for this variable
- This suburb does not border the river
- This suburb has a relatively high NOx concentration, above the 3rd quartile
- This suburb has, on average, 5.3 rooms per home, which is just under the 1st quartile mark
- 100% of homes were built prior to 1940
- This suburb is relatively close to Boston employment centres, as measured by the weighted mean of distances to 5 different centres.
- This suburb has the highest index of accessibility to radial highways
- The tax rate per $10k is 666, at exactly the 3rd quartile mark
- The pupil-teacher ratio, at 20, is relatively high, indicating that there are fewer teachers as compared to other suburbs
- This suburb has the highest proportion of black residents
- This suburb has a high proportion of the population classified as lower status, and falls in the fourth quartile of this distribution

Overall, these statistics suggest that suburb 399 is a poorer suburb, with a very dense population, high pollution, and substantially antiquated infrastructure.