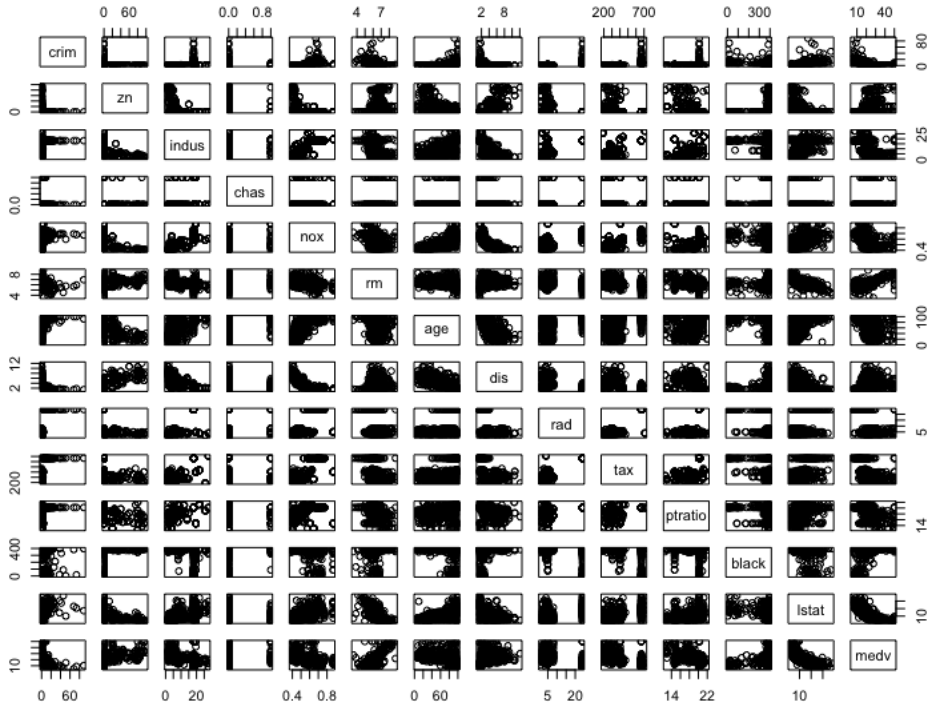
Abdoulaye Niamou
Data Mining Homework1
1.
Looking at the cereal data, the first 3 columns are names and types of processes that were non-numerical and therefore were removed keeping only numerical data. The next step to cleansing my data was looking at histograms in order to visually identify outliers. Also, with so many variables, creating a scatter plot of everything at once would make it difficult to identify trends and outliers as shown below.



As an alternative, the stem and leaf plot was used a stem and leaf plot to outline the outliers in each variable. For example:

results of stem and leaf plots
1 | 8
2 | 0223478899
3 | 000011123444566667789999
4 | 000011111224566779
5 | 00112233355899
6 | 013588
7 | 34
8 |
9 | 4

This is the result of the stem and leaf plot from the rating data. Outliers are defined by any result that contains less than 3 elements. In the case of the stem and leave plots above, the rows

containing data with 18, 73,74, and 94 were considered outliers and removed. The process was repeated for every variable. Any missing values were also removed.

2A
significant
calories, protein, fat, sodim, fibers, carbo, sugar, potass, vitamins,

Residuals:

Min	1Q	Median	3Q	Max
-5.243e-07	-2.577e-07	4.643e-08	2.264e-07	5.657e-07

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.493e+01	3.630e-07	1.513e+08	<2e-16	***
calories	-2.227e-01	5.663e-09	-3.933e+07	<2e-16	***
protein	3.273e+00	5.092e-08	6.428e+07	<2e-16	***
fat	-1.691e+00	6.226e-08	-2.717e+07	<2e-16	***
sodium	-5.449e-02	4.962e-10	-1.098e+08	<2e-16	***
fiber	3.443e+00	4.309e-08	7.992e+07	<2e-16	***
carbo	1.092e+00	1.743e-08	6.268e+07	<2e-16	***
sugars	-7.249e-01	1.819e-08	-3.986e+07	<2e-16	***
potass	-3.399e-02	1.473e-09	-2.307e+07	<2e-16	***
vitamins	-5.121e-02	1.928e-09	-2.657e+07	<2e-16	***
shelf	-3.721e-08	5.285e-08	-7.040e-01	0.484	
weight	-4.298e-07	5.206e-07	-8.260e-01	0.412	
cups	1.379e-07	1.924e-07	7.170e-01	0.476	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

As show by the model summary above, the predictors which appear to have a significant relationship to the response arecalories, protein, fat, sodim, fibers, carbo, sugar, potass, and vitamins. They have the smallest p-values.

2B
When every other predictor held constant, the mpg value increases with each year that passes. Specifically, mpg increase by 1.43 each year.-7.249e-01

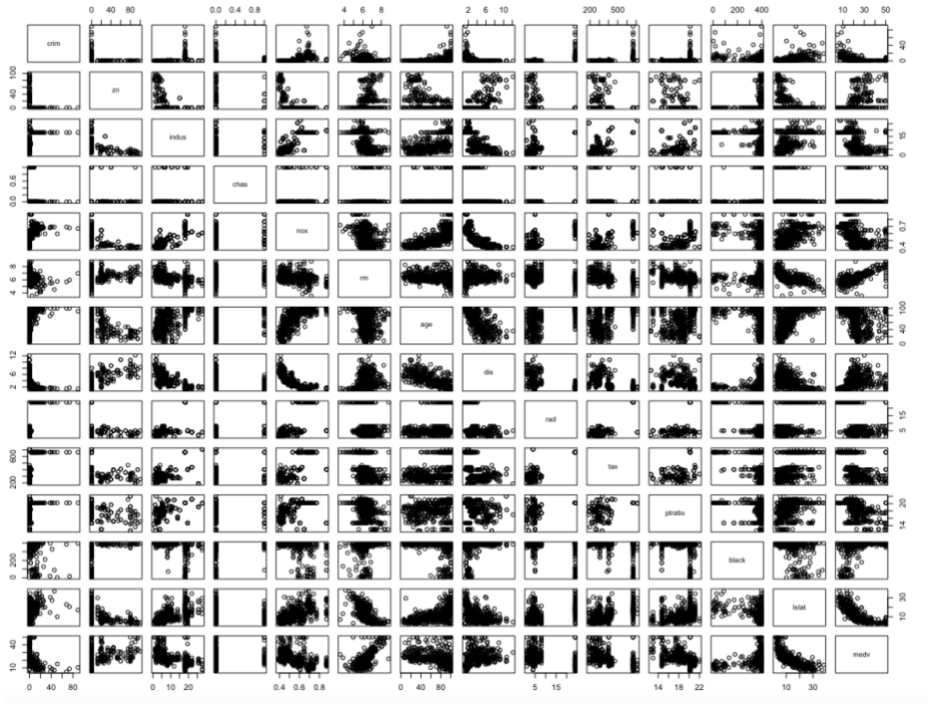
3.

Classifier	Training error	Test Error
Linear regression	0.58%	4.12%
k=1	0.00 %	2.50%
k=3	0.50%	3.02%
k=5	0.58%	3.02%
k=7	0.65%	3.30%
k=9	0.94%	3.57%
k=11	0.86%	3.57%
k=13	0.86%	3.85%
k=15	0.94%	3.85%

In this case, we can see that KNN performs better than linear regression. For the K-nearest neighbor however, the training error as well as the Test error increases as k increases. Therefore K = 1 has the lowest test error

4.

(a) With so many variables, it is hard to tell if some there is some correlation, however some do have areas where the data is dense and others where the data seems completely scattered.



(b)

Based on the correlation coefficients shown above, it seems that the capita crima rate a significant correlations with the other predictors.

```
> cor(Boston$crim,Boston$zn)
[1] -0.2004692
> cor(Boston$crim,Boston$indus)
[1] 0.4065834
> cor(Boston$crim,Boston$chas)
[1] -0.05589158
> cor(Boston$crim,Boston$nox)
[1] 0.4209717
> cor(Boston$crim,Boston$rm)
[1] -0.2192467
> cor(Boston$crim,Boston$age)
[1] 0.3527343
> cor(Boston$crim,Boston$dis)
[1] -0.3796701
> cor(Boston$crim,Boston$rad)
[1] 0.6255051
> cor(Boston$crim,Boston$tax)
[1] 0.5827643
> cor(Boston$crim,Boston$ptratio)
[1] 0.2899456
> cor(Boston$crim,Boston$black)
[1] -0.3850639
> cor(Boston$crim,Boston$lstat)
[1] 0.4556215
> cor(Boston$crim,Boston$medv)
[1] -0.3883046
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

The significant variables are “zn”, “dis”,”rad”, “black”, and “medv” .

(c)

d)

there are 64 suburbs with more than 7 rooms per dwelling and 13 with more than 8 rooms per dwelling.

```
###Question D
rm_more_than_7 <- subset(Boston, rm>7)
nrow(rm_more_than_7)
### [1] 64


rm_more_than_8 <- subset(Boston, rm>8)
nrow(rm_more_than_8)
###[1] 13
```