The honor code

(a) Please state the names of people who you worked with for this homework. You can also provide your comments about the homework here.

Cinidy Liu

(b) Please type/write the following sentences yourself and sign at the end. We want to make it *extra* clear that nobody cheats even unintentionally.

I hereby state that all of my solutions were entirely in my words and were written by me. I have not looked at another student's solutions and I have fairly credited all external sources in this write up.

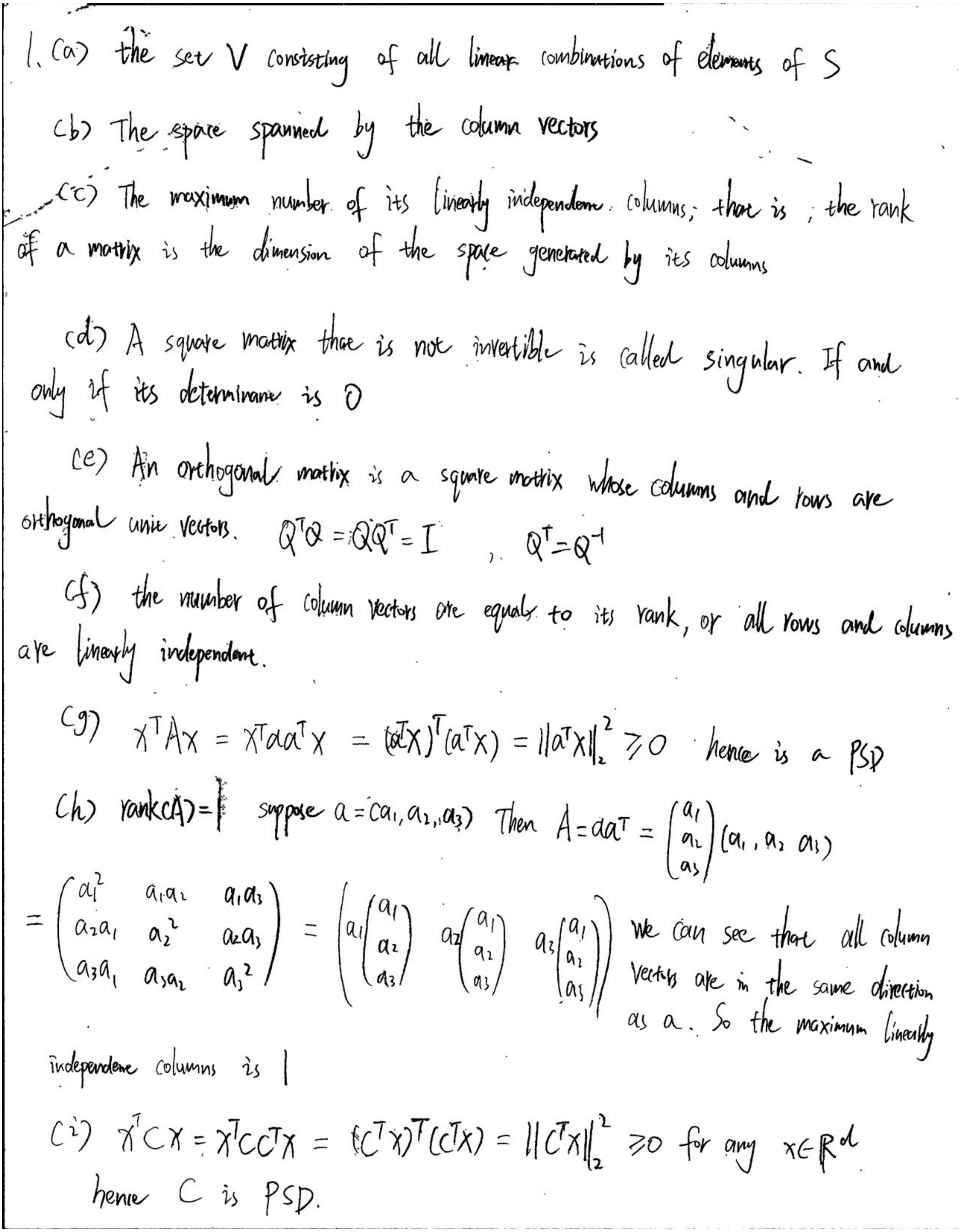
I hereby state that all of my solutions were entirely in my words and were written by me. I have not looked at another student's solutions and I have fairly credited all external sources in this write up.

hw1

caojilin

1/25/2019

1



1. (a) the set V consisting of all linear combinations of elements of S

(b) The space spanned by the column vectors

(c) The maximum number of its linearly independent columns; that is, the rank of a matrix is the dimension of the space generated by its columns

(d) A square matrix that is not invertible is called singular. If and only if its determinane is 0

(e) An orthogonal matrix is a square matrix whose columns and rows are orthogonal unit vectors. $Q^TQ = QQ^T = I$, $Q^T = Q^{-1}$

(f) the number of column vectors are equals to its rank, or all rows and columns are linearly independant.

(g) $x^TAx = x^Taa^Tx = (a^Tx)^T(a^Tx) = \|a^Tx\|_2^2 \geqslant 0$ hence is a PSD

(h) rank(A) = 1 suppose $a = (a_1, a_2, a_3)$ Then $A = aa^T = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} (a_1, a_2\ a_3)$

$$= \begin{pmatrix} a_1^2 & a_1a_2 & a_1a_3 \\ a_2a_1 & a_2^2 & a_2a_3 \\ a_3a_1 & a_3a_2 & a_3^2 \end{pmatrix} = \left(a_1\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad a_2\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad a_3\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \right)$$

We can see that all column vectors are in the same direction as a. So the maximum linearly independent columns is 1

(i) $x^TCx = x^TCC^Tx = (C^Tx)^T(C^Tx) = \|C^Tx\|_2^2 \geqslant 0$ for any $x \in \mathbb{R}^d$

hence C is PSD.

2 Eigendecomposition with R

a.

```
X = matrix(c(1,2,2,1),2,2,byrow = TRUE)
Y = matrix(c(-1/3,2/3,2/3,-1/3),2,2,byrow = TRUE)
X
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    2    1
```

```
Y
```

```
##            [,1]       [,2]
## [1,] -0.3333333  0.6666667
## [2,]  0.6666667 -0.3333333
```

b.

```
solve(X)
```

```
##            [,1]       [,2]
## [1,] -0.3333333  0.6666667
## [2,]  0.6666667 -0.3333333
```

```
solve(Y)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    2    1
```

both X and Y are invertible because their determinants are not zero. One could use `solve` function

c.

```
eigen(X)
```

```
## eigen() decomposition
## $values
## [1]  3 -1
## 
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

d. Since Y is the inverse of X, it has eigenvalues $\lambda' = \frac{1}{\lambda}$. So Y have the same eigenvectors as X and 1/3 and -1 as eigenvalues.

$$A^{-1}Ax = A^{-1}\lambda x$$
$$x = \lambda A^{-1}x$$
$$A^{-1}x = \frac{1}{\lambda}x$$

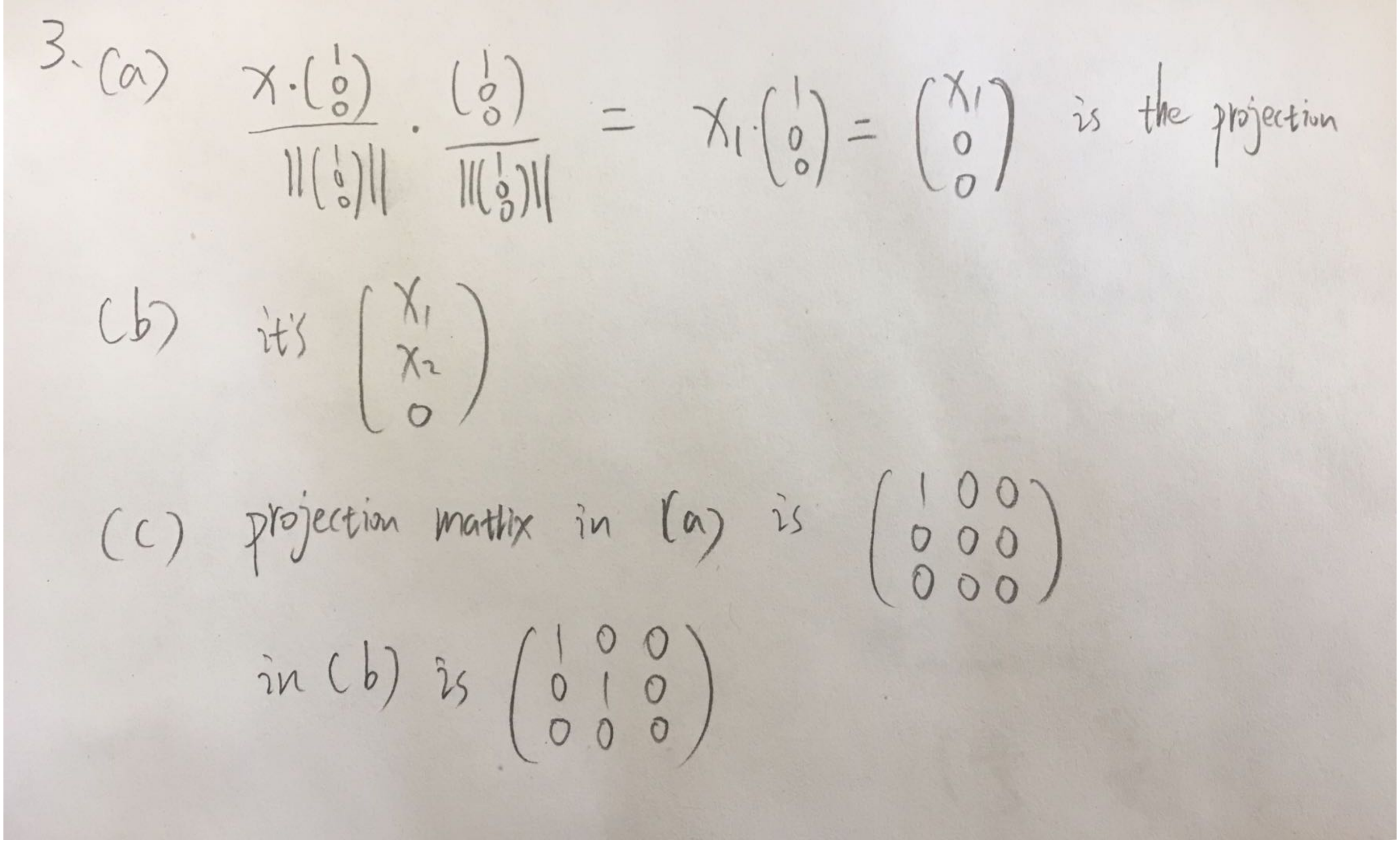
e. A^2 has the square of A’s eigenvalues and same eigenvectors. So eigenvalues of X^2 are 1 and 9.

$$\begin{aligned}A^2x &= A\lambda x\\ &= \lambda Ax\\ &= \lambda^2 x\end{aligned}$$

f.

$$Ax = \lambda x$$
$$A^2x = A\lambda x$$
$$A^2x = \lambda Ax$$
$$A^2x = \lambda\lambda x$$
$$A^2x = \lambda^2 x$$

3 Understanding orthogonal projection



3. (a) $\frac{x\cdot\begin{pmatrix}1\\0\\0\end{pmatrix}}{\left\|\begin{pmatrix}1\\0\\0\end{pmatrix}\right\|}\cdot\frac{\begin{pmatrix}1\\0\\0\end{pmatrix}}{\left\|\begin{pmatrix}1\\0\\0\end{pmatrix}\right\|} = x_1\cdot\begin{pmatrix}1\\0\\0\end{pmatrix} = \begin{pmatrix}x_1\\0\\0\end{pmatrix}$ is the projection

(b) it's $\begin{pmatrix}x_1\\x_2\\0\end{pmatrix}$

(c) projection matlix in (a) is $\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}$

in (b) is $\begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix}$

d.

```
a = matrix(c(1,rep(0,8)),3,3,byrow = TRUE)
b = matrix(c(1,0,0,0,1,0,0,0,0),3,3,byrow = TRUE)
set.seed(123)
vec1 = runif(3)
vec2 = runif(3)
a %*% vec1
```

```
##           [,1]
## [1,] 0.2875775
## [2,] 0.0000000
## [3,] 0.0000000
```

```
a %*% vec2
```

```
##           [,1]
## [1,] 0.8830174
## [2,] 0.0000000
## [3,] 0.0000000
```

```
b %*% vec1
```

```
##           [,1]
## [1,] 0.2875775
## [2,] 0.7883051
## [3,] 0.0000000
```

```
b %*% vec2
```

```
##           [,1]
## [1,] 0.8830174
## [2,] 0.9404673
## [3,] 0.0000000
```

e. the projetion matrix is $P = X(X^TX)^{-1}X^T$ or let the unit vector $u = \frac{a}{||a||}$ and $P = uu^T$

f.

```
projection = function(a,x){
  P =a %*% solve(t(a) %*% a) %*% t(a)
  return(P %*% x)
}
x = c(3,2,-1)
a = c(1,0,1)
projection(a,x)
```

```
##      [,1]
## [1,]    1
## [2,]    0
## [3,]    1
```

```
au = a/sqrt(2)#unit vector
pp = t(t(au)) %*% t(au)
pp %*% x
```

```
##      [,1]
## [1,]    1
## [2,]    0
## [3,]    1
```

g. let X=(a1,a2), orthogonal projection = $X(X^TX)^{-1}X^Tx$ or, the vector addition of orthogonal projection of x onto a1 and orthogonal projection of x onto a2

h. use Gram Schmidt Orthogonalization to convert a1 and a2 into orthonormal basis u1, u2. Let v1 be the projection of x onto u1, and Let v2 be the projection of x onto u2 Then orthogonal projection x onto span(a1,a2) is v1+v2
or just use projection matrix $P = X(X^TX)^{-1}X^T$

i.

```
# Gram Schmidt
u = gramSchmidt(X)
u1  = u$Q[,1]
u2 =  u$Q[,2]
v1 = project(x, u1)
v2 = project(x, u2)
v1 + v2
```

```
## [1] 1.444444 1.777778 1.444444
```

j. use Gram Schmidt Orthogonalization to convert a1,a2,…,ak into orthonormal basis u1,u2,…,uk. Let the projection of x onto u1 be v1, the projection of x onto u2 be v2 and so on. The final result is v1+v2+…+vk

l.

```
X = matrix(c(1,0,1,1,-1,0),3,2)
P =X %*% solve(t(X) %*% X) %*% t(X)
P %*% x
```

```
##               [,1]
## [1,]  1.000000e+00
## [2,] -5.551115e-17
## [3,]  1.000000e+00
```

We got the same result as in part (i)

m. Column vectors are orthogonal. \$A(A$^{\{T\}A)}${-1}A^{T}x = A(D)A^{T}x \$ where D is a diagnoal matrix with $a_i^T a_i$ on the diagonal.

4 Exploring a dataset with R

a.

```
library(MASS)
?Boston
nrow(Boston)
```

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

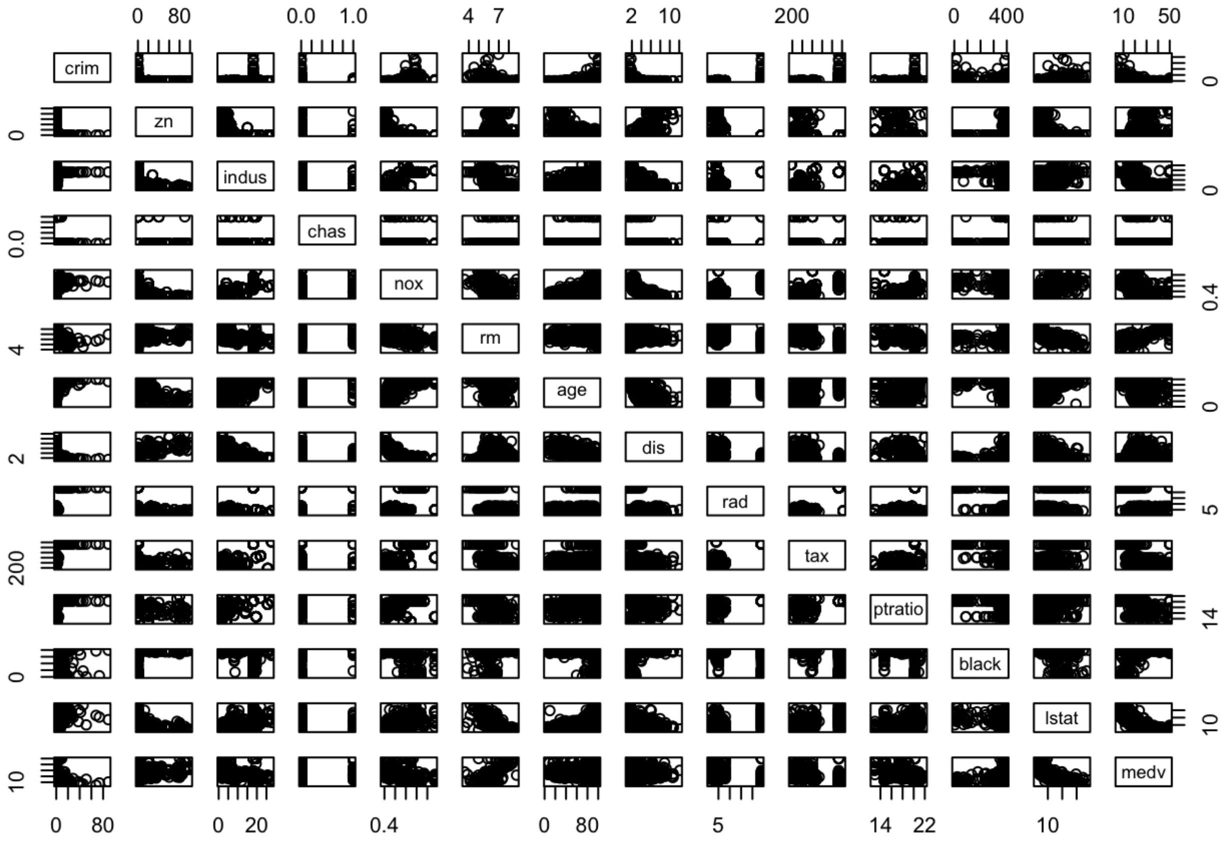
```
ncol(Boston)
```

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

The data consists of 14 variables (i.e. features) and 506 observations. Each observation is a subdivision of a county in Boston. The 14 variables are:
1. crim: per capita crime rate by town
2. zn: proportion of residential land zoned for lots over 25,000 sq.ft
3. indus: proportion of non-retail business acres per town
4. chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. nox: nitric oxides concentration (parts per 10 million)
6. rm: average number of rooms per dwelling
7. age: proportion of owner-occupied units built prior to 1940
8. dis: weighted distances to five Boston employment centres
9. rad: index of accessibility to radial highways
10. tax: full-value property-tax rate per USD 10,000
11. ptratio: pupil-teacher ratio by town
12. b: 1000(B - 0.63)^2 where B is the proportion of blacks by town
13. lstat: percentage of lower status of the population
14. medv: median value of owner occupied homes in USD 1000’s

b.

```
pairs(Boston)
```



```
# library(GGally)
# ggpairs(Boston)
```

There are some variables having positive relationship like `medv` and `rm`. and negative relationship like `medv` and `lstat`.

c. the less the dis variable, the more crime rates.
   the less the zn variable, the more crime rates.
   the more the rad variable, the more crime rates.

d.

```
#There are many of the suburbs of Boston appear to have particularly high crime rates, t
ax rates and Pupil-teacher ratios. Below are the ranges for each variable.
range(Boston$crim)
```

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

```
range(Boston$tax)
```

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

```
range(Boston$ptratio)
```

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

e.

```
nrow(Boston[Boston$chas == 1,])
```

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

f.

```
median(Boston$ptratio)
```

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

g.

```
Boston[Boston$medv == min(Boston$medv),]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97
##     lstat medv
## 399 30.59    5
## 406 22.98    5
```

```
#Age is at the max value. Tax is close to the max 711. Ptratio is close to the max 22.0
#the crime rates are different for two suburbs.
```

h.

```
#more than seven
nrow(Boston[Boston$rm >7,])
```

```
## [1] 64
```

```
#more than eight
nrow(Boston[Boston$rm >8,])
```

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

```
Boston[Boston$rm >8,]
```

```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio  black
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55
##     lstat medv
## 98   4.21 38.7
## 164  3.32 50.0
## 205  2.88 50.0
## 225  4.14 44.8
## 226  4.63 50.0
## 227  3.13 37.6
## 233  2.47 41.7
## 234  3.95 48.3
## 254  3.54 42.8
## 258  5.12 50.0
## 263  5.91 48.8
## 268  7.44 50.0
## 365  5.29 21.9
```

```
#The age variable is big, and the medv variable is also big on average.
```

i. all the suburbs in Boston

j. Reality: we have BostonHousing data with 506 observations and 14 variables. Model: linear regression Future Reality: predict the median housing price for a new suburb in Boston.

5 True or false

a. **Cross validation is a powerful tool to select hyper-parameters in several machine learning tasks.** False It’s a method for selecting model not hyper-parameters.

b. **Cross validation error is always a good proxy for the prediction error.** False not always. Cross validation error uses hold out data set, but the true prediction error uses “future” test set. They are not same.

c. **Vanilla cross validation is a good idea for time-series data.** False. time-series data has correlations between data based on time. The k-fold hold out data set in time-series data is not exchangeable. The CV error is not a good index.

d. **For a machine learning problem, exploratory data analysis by itself is generally sufficient to determine the complete relevance of the dataset for the problem.** False. There are relevance in data we cannot observe by EDA only. EDA can give us a basic sense but we need more method and model to dig out more relevance in the dataset.

e. **Data collection process usually has no-to-little influence on the outcome of a prediction problem.**
False. Data collection determines a lot on how we formulate the problem and what the representives of the population are.

f. **For a model to make meaningful predictions on the future data, we need some similarity between the representative data that was used to build the model and the future data.**
True. It’s true that we need some similarity. For example, we cannot use US presidential election data to predict Russia presidential election.

g. **Prediction is often the end goal of a machine learning task.**
False. Prediction sometimes is not enough. In order to find out the meaning behind the model we need to do more work. For example, how do we explain the relationship between smoking and lung canser. Can we conclude causation based on correlation and prediction performance?