Multi. Stat. HW1

赵浩宇 2016012390

1 Problem 1

If we let the distribution to be $N(0,1)$, the percentage the the data that lies outside the outside bars is about to be 0.7%.

Proof.

$$Q_3 = \Phi^{-1}(0.75) \approx 0.6745$$
$$Q_1 = -Q_3 \approx -0.6745$$
$$IQR = Q_3 - Q_1 \approx 1.349$$

$$\begin{aligned} upper\_outlier &= Q_3 + 1.5 \cdot IQR \\ &= 0.6745 + 1.5 \times 1.349 \\ &= 2.698 \\ &\approx 2.7 \end{aligned}$$

So the portion that is bigger than the upper outside bar is about

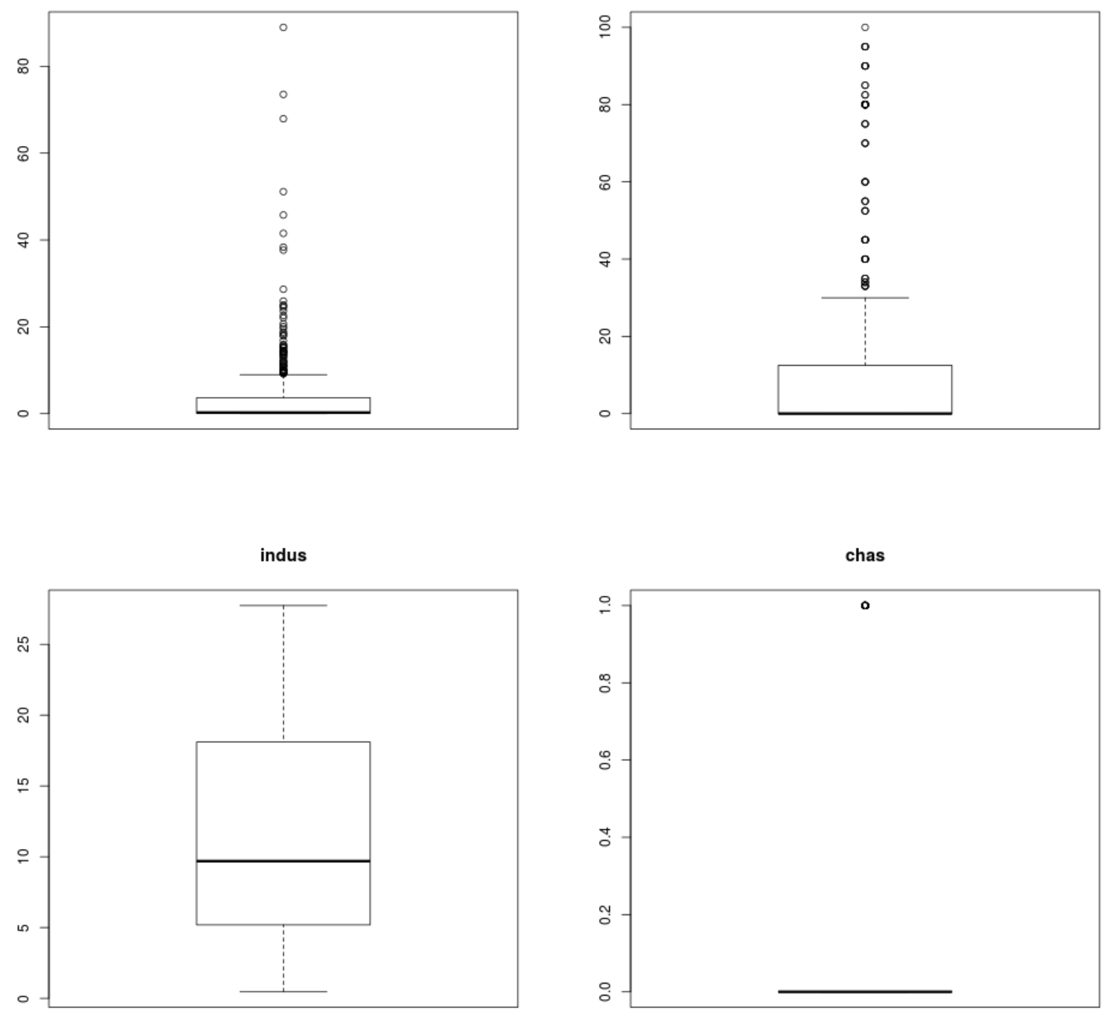
$$1 - \Phi^{-1}(2.7) \approx 0.0035$$

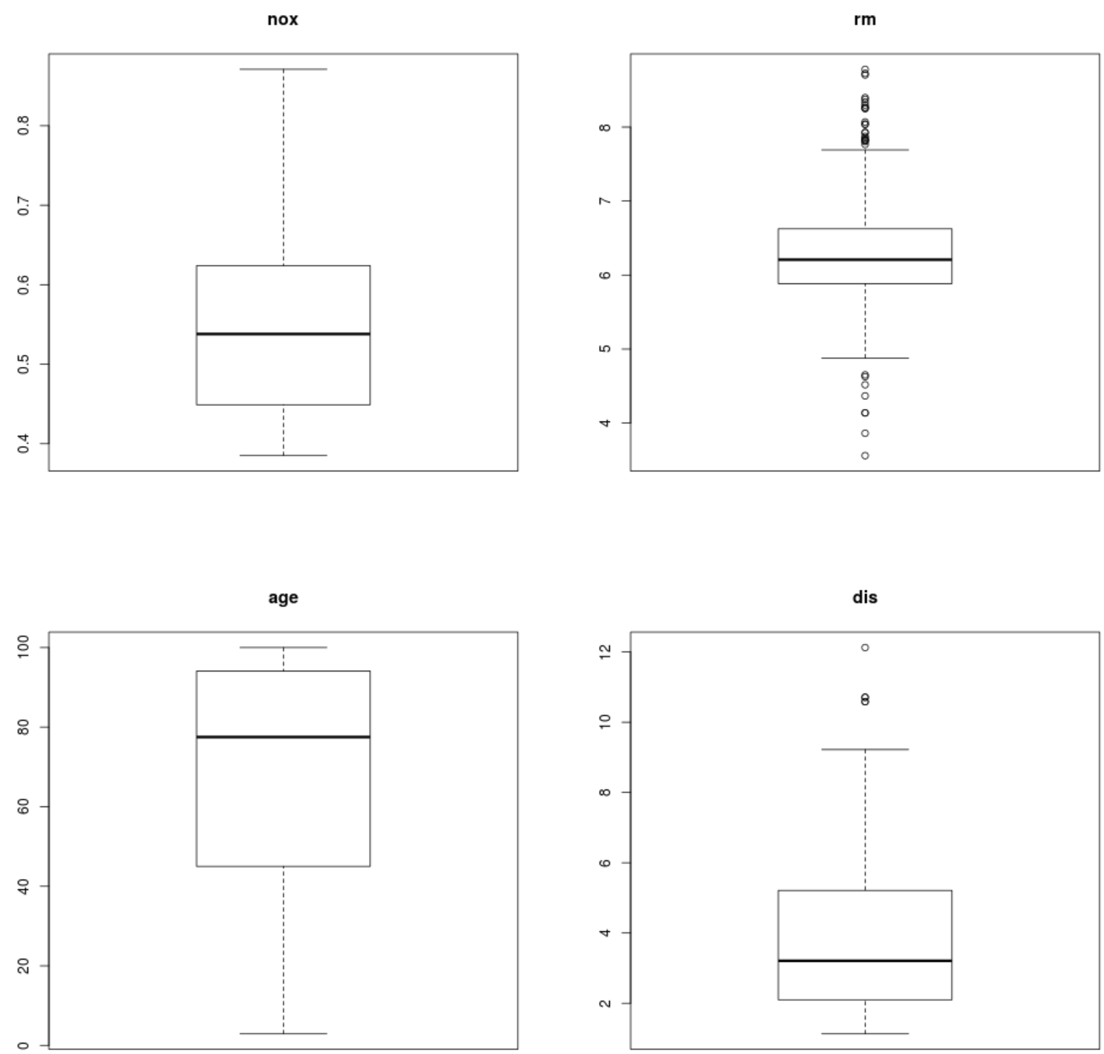
So the portion that lie outside the outside bars should be about $0.007 \approx 0.7\%$. □

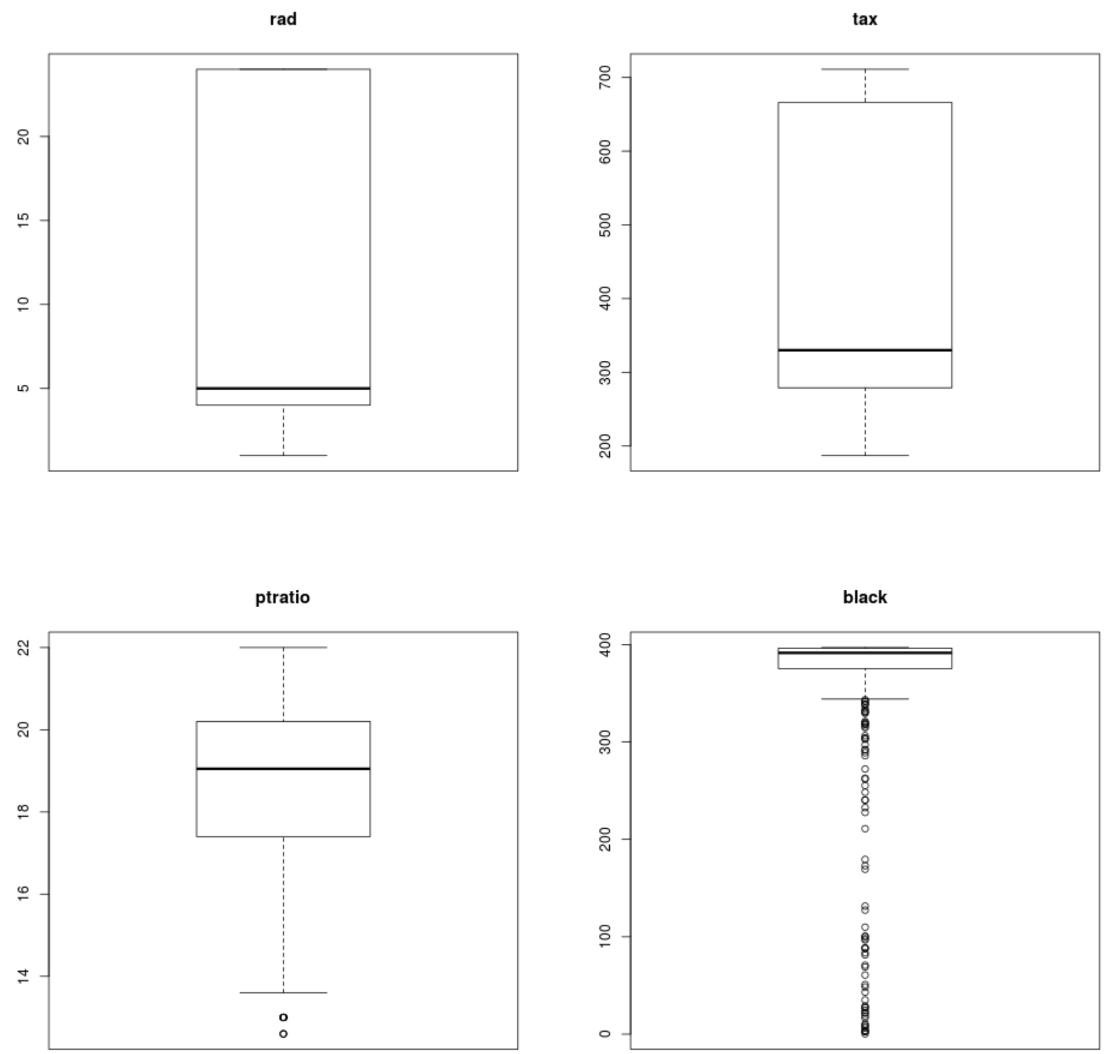
If the data follow the distribution $N(0, \sigma^2)$, then the portion should also be about 0.7%.
We can get the data follow the distribution $N(0, \sigma^2)$ by multiplying the standard normal distribution by σ, and the quantiles, IQR, and the value of the outside bars are also multiplied by σ. The data follows $N(0, \sigma^2)$ that lies outside the outside bars also lies outside the outisde bars (of the dist. $N(0,1)$) when it is devided by σ, so the percentage of data that lies outside the outside bars in this problem is the same as the previous problem.

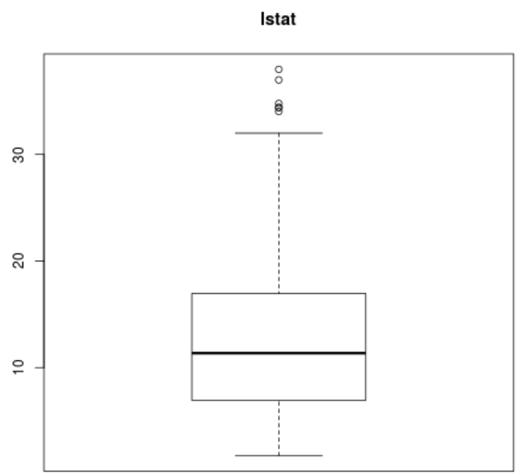
2 Problem 2

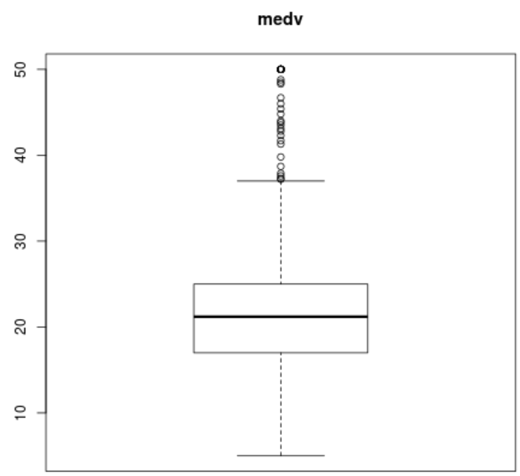
a.



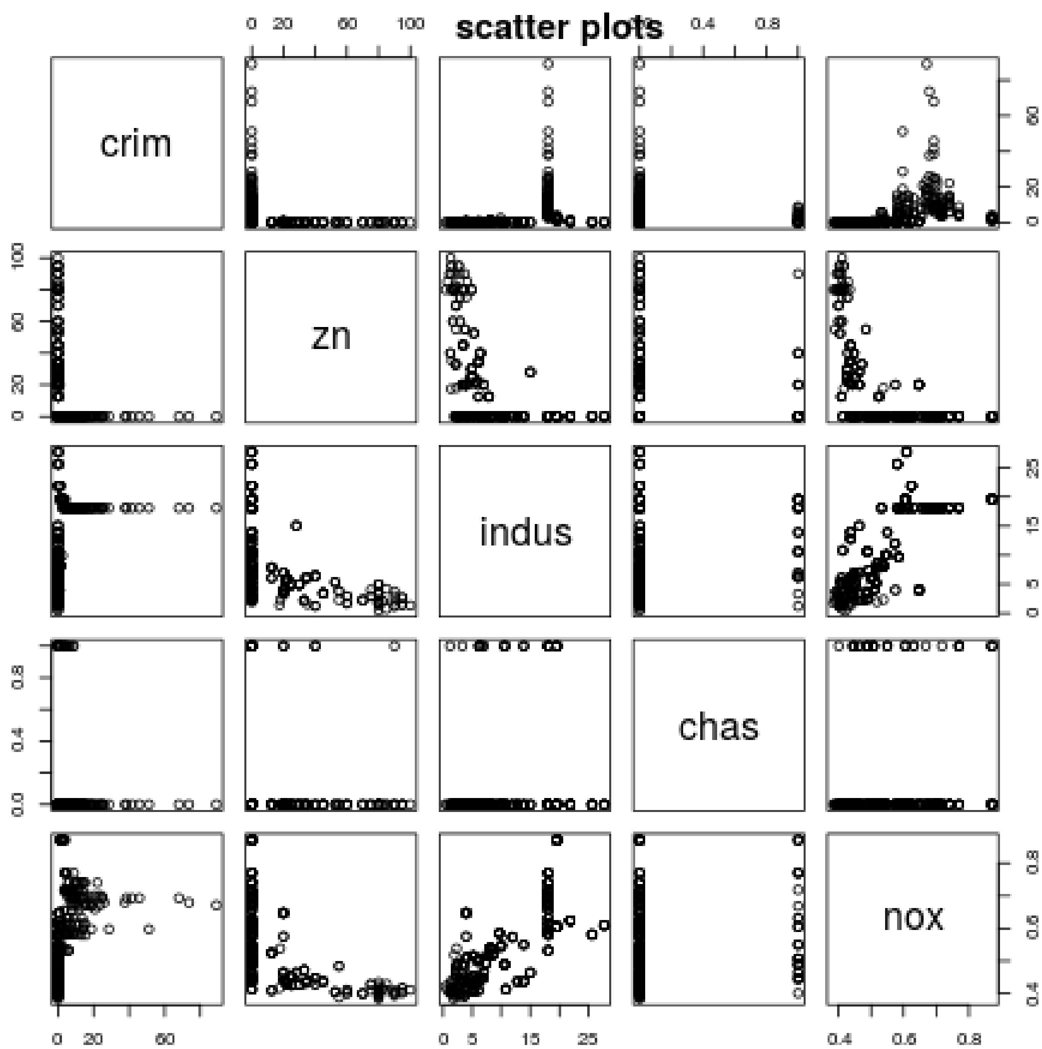

nox
0.4
0.5
0.6
0.7
0.8
rm
4
5
6
7
8
age
0
20
40
60
80
100
dis
2
4
6
8
10
12


rad
tax
ptratio
black


lstat
10
20
30


medv
10
20
30
40
50

b.



The figure above is the matrix scatter plots for the first five variables.

c.

The correlation matrix with 4 digits is showned below.

```
> round(cor(data),4)
           crim      zn   indus    chas     nox      rm     age     dis     rad     tax ptratio   black   lstat    medv
crim     1.0000 -0.2005  0.4066 -0.0559  0.4210 -0.2192  0.3527 -0.3797  0.6255  0.5828  0.2899 -0.3851  0.4556 -0.3883
zn      -0.2005  1.0000 -0.5338 -0.0427 -0.5166  0.3120 -0.5695  0.6644 -0.3119 -0.3146 -0.3917  0.1755 -0.4130  0.3604
indus    0.4066 -0.5338  1.0000  0.0629  0.7637 -0.3917  0.6448 -0.7080  0.5951  0.7208  0.3832 -0.3570  0.6038 -0.4837
chas    -0.0559 -0.0427  0.0629  1.0000  0.0912  0.0913  0.0865 -0.0992 -0.0074 -0.0356 -0.1215  0.0488 -0.0539  0.1753
nox      0.4210 -0.5166  0.7637  0.0912  1.0000 -0.3022  0.7315 -0.7692  0.6114  0.6680  0.1889 -0.3801  0.5909 -0.4273
rm      -0.2192  0.3120 -0.3917  0.0913 -0.3022  1.0000 -0.2403  0.2052 -0.2098 -0.2920 -0.3555  0.1281 -0.6138  0.6954
age      0.3527 -0.5695  0.6448  0.0865  0.7315 -0.2403  1.0000 -0.7479  0.4560  0.5065  0.2615 -0.2735  0.6023 -0.3770
dis     -0.3797  0.6644 -0.7080 -0.0992 -0.7692  0.2052 -0.7479  1.0000 -0.4946 -0.5344 -0.2325  0.2915 -0.4970  0.2499
rad      0.6255 -0.3119  0.5951 -0.0074  0.6114 -0.2098  0.4560 -0.4946  1.0000  0.9102  0.4647 -0.4444  0.4887 -0.3816
tax      0.5828 -0.3146  0.7208 -0.0356  0.6680 -0.2920  0.5065 -0.5344  0.9102  1.0000  0.4609 -0.4418  0.5440 -0.4685
ptratio  0.2899 -0.3917  0.3832 -0.1215  0.1889 -0.3555  0.2615 -0.2325  0.4647  0.4609  1.0000 -0.1774  0.3740 -0.5078
black   -0.3851  0.1755 -0.3570  0.0488 -0.3801  0.1281 -0.2735  0.2915 -0.4444 -0.4418 -0.1774  1.0000 -0.3661  0.3335
lstat    0.4556 -0.4130  0.6038 -0.0539  0.5909 -0.6138  0.6023 -0.4970  0.4887  0.5440  0.3740 -0.3661  1.0000 -0.7377
medv    -0.3883  0.3604 -0.4837  0.1753 -0.4273  0.6954 -0.3770  0.2499 -0.3816 -0.4685 -0.5078  0.3335 -0.7377  1.0000
> 
```

3 Problem 3

Proof.

$$r_{xy} = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{s_x \cdot s_y}$$

According to the conditions, the data was changed linearly, we can get:

$$\begin{aligned}
s_i &= a x_i + b \\
t_i &= c y_i + d \\
\overline{s} &= a\overline{x} + b \\
\overline{t} &= c\overline{y} + d \\
s_s &= a s_x \\
s_t &= c s_y
\end{aligned}$$

Then we can get:

$$\begin{aligned}
r_{st} &= \frac{\sum_{i=1}^{n} \left(s_i - \overline{s}\right)\left(t_i - \overline{t}\right)}{n \cdot s_s \cdot s_t} \\
&= \frac{\sum_{i=1}^{n} \left(a x_i + b - a\overline{x} - b\right)\left(c y_i + d - c\overline{y} - d\right)}{n \cdot a \cdot s_x \cdot c \cdot s_y} \\
&= \frac{1}{n} \frac{\sum_{i=1}^{n} a \cdot \left(x_i - \overline{x}\right) \cdot c \cdot \left(y_i - \overline{y}\right)}{a \cdot s_x \cdot c \cdot s_y} \\
&= \frac{1}{n} \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{s_x \cdot s_y} \\
&= r_{xy}
\end{aligned}$$

So the linear transformation does not change the sample correlation. $\square$

Problem 4

2.8

The characteristic polynomial of the matrix is

$$f(\lambda) = (1-\lambda)(-2-\lambda) - 2^2.$$

Set $f(\lambda) = 0$ and we can get the root of $f(\lambda)$,

$$\lambda_1 = 2, \lambda_2 = -3.$$

Let

$$\mathbf{e}_1 = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{bmatrix}$$

We can see that

$$\mathbf{A}\mathbf{e}_1 = \lambda_1 \mathbf{e}_1, \mathbf{A}\mathbf{e}_2 = \lambda_2 \mathbf{e}_2, \|\mathbf{e}_1\| = \|\mathbf{e}_2\| = 1,$$

So $\mathbf{e}_1$ and $\mathbf{e}_2$ are 2 normalized eigenvectors of $\mathbf{A}$. And according to the sqectual theorem,

$$A = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T = 2 \begin{bmatrix} \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \end{bmatrix} - 3 \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} - \frac{2}{\sqrt{5}} \end{bmatrix}.$$

2.9

a.

$$B = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{6} \end{bmatrix}$$

Then we can find that $AB = BA = I$, so $A_{-1} = B$.

b.

$$S = [\mathbf{e}_1, \mathbf{e}_2], \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix}, \Sigma^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{3} \end{bmatrix}.$$

Then we have $A = S\Sigma S^{-1}$. Because $S\Sigma S^{-1} S\Sigma^{-1} S^{-1} = S\Sigma\Sigma^{-1} S^{-1} = I$, we can know that $A^{-1} = S\Sigma^{-1} S^{-1}$, so the eigenvalues of A^{-1} is,

$$\mu_1 = \frac{1}{2}, \mu_2 = -\frac{1}{3}.$$

And the normalized eigenvectors remain the same as the eigenvectors of A, so

$$\mathbf{v}_1 =$$

c.

2.15

The quadratic form

$$\mathbf{Q} = \mathbf{x}^T \mathbf{A} \mathbf{x}, \mathbf{x}^T = [x_1, x_2], \mathbf{A} = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

We can compute the eigenvalues of $\mathbf{A}$,

$$\lambda_1 = 4, \lambda_2 = 2.$$

So it's positive definite.