Problem 2.

For this question, I did the calculation by hand.

$\pi_0 = [$ Not Drive, Not Drive, Not Drive $]$ (t = Top, r = rolling, b = bottom)

Top Rolling bottom

First round:

$$\begin{cases} b = 1 + 0.8b \\ r = 1 + 0.8b \\ t = 3 + 0.8 \cdot (0.3r + 0.7t) \end{cases} \Rightarrow \begin{cases} r = 5 \\ b = 5 \\ t = 9.55 \end{cases}$$

$\pi_1(\text{Top}) = \arg\max_a [9.276, 9.54] =$ Not Drive (Frist index is Drive, Second index is not Drive)

$\pi_1(\text{Rolling}) = \arg\max_a [5.09, 5.0] =$ Drive

$\pi_1(\text{bottom}) = \arg\max_a [6.18, 5.0] =$ Drive

$\pi_1 = [$ Not Drive, Drive, Drive $]$

Second round:

$$\begin{cases} t = 3 + 0.8(0.7t + 0.3r) \\ b = 0.8(0.6t + 0.4b) \\ r = (0.3t + 0.6r + 0.1b) \times 0.8 \end{cases} \Rightarrow \begin{cases} r = 5.64 \\ t = 9.89 \\ b = 6.99 \end{cases}$$

$\pi_1(\text{Top}) = \arg\max_a [9.57, 9.89] =$ Not Drive

$\pi_1(\text{Roling}) = \arg\max_a [5.64, 6.60] =$ Not Drive $\Rightarrow \pi_2 = [$Not Drive, Not Drive, Drive$]$

$\pi_1(\text{bottom}) = \arg\max_a [6.98, 6.59] =$ Drive

Third round:

$$\begin{cases} t = 3 + 0.8(0.3R + 0.7T) \\ r = 0.8b + 1 \\ b = (0.6t + 0.4b)\, 0.8 \end{cases} \Rightarrow \begin{cases} r = 7.01 \\ t = 10.64 \\ b = 7.51 \end{cases}$$

$\pi_2(\text{Top}) = \arg\max_a [10.22, 10.64] =$ Not Drive

$\pi_2(\text{Rolling}) = \arg\max_a [6.52, 7.01] =$ Not Drive

$\pi_3(\text{bottom}) = \arg\max_a [7.51, 7.01] =$ Drive

$\pi_3 = [$ Not Drive, Not Drive, Drive $]$. $\longrightarrow$ optimal policy.

As a result, $\pi_2 = \pi_3$, the policy converges after the thrid literation.

The optimal Value is [10.64, 7.01, 7.51] which is the same as question 2.

I use some of the code in q_1 to help me caculating π since the second part of the algorithm is the same except solving for equations.