Question 1

```
Question 1a
Reward =  10
Values
54.60937951381121 55.04911731288497 10
55.09020008304064 54.5553632789769 55.04911731288497
54.54242981429825 55.09020008304064 54.60937951381121
Reward =  0
Values
9.227980187276401 9.77011948906488 10
8.6459691684564 9.303139788528961 9.77011948906488
7.082489751266161 8.6459691684564 9.227980187276401
Reward =  -100
Values
-255.49944021774715 -127.87116523265007 10
-341.83874510385834 -243.60779146620018 -127.87116523265007
-367.6109593540774 -341.8387451038583 -255.49944021774715
Question 1b
Values
6.272640000000001 8.704080000000001 10
0.0 1.5681600000000002 1.77408
0.0 0.0 0.09801000000000001
Policy
State:  (0,0) , Action:  up
State:  (1,0) , Action:  up
State:  (2,0) , Action:  up
State:  (0,1) , Action:  up
State:  (1,1) , Action:  up
State:  (2,1) , Action:  up
State:  (0,2) , Action:  right
State:  (1,2) , Action:  right
State:  (2,2) , Action:  exit
```

For reward = 10, the terminal state gives less reward than staying alive, except for the first move where you can get 10 on the transition plus 10 for reaching the goal state, so states that can avoid approaching it have the greatest value as time goes on.
For reward = 0, the terminal state gives the only reward, so value is based on how long it takes to get there.
For reward = -100, the above is also true, but you also get a large negative reward for taking a long time to reach the terminal state.

Question 2

a)

Start	action	end	reward	count	T	R
A	exit	x	-5	2	1	-5
B	east	C	-1	2	1	-1
C	east	A	-1	2	0.5	-1
C	east	D	-1	2	0.5	-1
D	exit	x	5	2	1	5
E	north	C	-1	2	1	-1

b)

Direct Evaluation	Equation	Value
A	(-7+-7)/2	-7
B	(-7+3)/2	-2
C	(-7+3+3+-7)/4	-2
D	(3+3)/2	3
E	(3+-7)/2	-2

c)

$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^{\pi}(s')\right]$$

V	A	B	C	D	E	V=a*V+a*(r+V')	
0	0	0	0	0	0		
1	0	1.5	0	0	0	VB=.5*VB+.5*(3+VC)	1.5=0+.5*(3+0)
2	0	1.5	1	0	0	VC=.5*VC+.5*(2+VE)	1=0+.5*(2+0)
3	0	1.5	2	0	0	VC=.5*VC+.5*(3+VE)	2=.5*1+.5*(3+0)
4	0	1.5	2	2	0	VD=.5*VD+.5*(2+VC)	2=0+.5*(2+2)
5	2.5	1.5	2	2	0	VA=.5*VA+.5*(3+VC)	2.5=0+.5*(3+2)

d)

$$sample = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\,[sample]$$

Q(s,a) = (1-alpha)*Q(s,a) + alpha * (R(s,a,s') + gamma * max a' (s', a'))
Q(B, East) = 0 + .5(3+0) = 1.5
Q(C, South) = 0 + .5(2 + 0) = 1
Q(C, East) = 0 + .5(3 + 0) = 1.5
Q(D, West) = 0 + .5(2 + 1.5) = 1.75
Q(A, South) = 0 + .5(3 + 1.5) = 2.25