Submitted by: Nazmus Saquib
CSE 574, Fall 2023
Assignment 3, UBIT: nsaquib2

Report for PART- 1

Ans to 1:

• Theme: Treasure hunt Grid World with diamonds as positive rewards and robber as negative rewards.

• States:

We have total 16 (4 x 4) States which are named as S1, S2,…., S15, S16. The definition of the states are given below with coordinates.

S1 = (0,0), S2 = (0,1), S3 = (0,2), S4 = (0,3), S5 = (1,0), S6 = (1,1), S7 = (1,2), S8 = (1,3), S9 = (2,0),

S10 = (2,1), S11 = (2,2), S12 = (2,3), S13 = (3,0), S14 = (3,1), S15 = (3,2), S16 = (3,3)}

• Actions: We have 4 actions to take which are Up, Down, Right and Left

• Rewards: We have 5 rewards such as -4, 0, +2, +6 and +9.

• Objective: Our objective is to reach the goal state (S16) with maximum reward.

Ans to 2:

Below is the visualization of 5 random steps of the agent.

```
Step No: 0
Current State: (0, 0), Action: 3 (Left), Reward: 0
[[ 0.  0.  0.  2.]
 [ 0.  2. -4.  0.]
 [ 0. -4.  0.  0.]
 [ 0.  6. -4.  9.]]
-----------------------------------------------------
Step No: 1
Current State: (0, 0), Action: 2 (Right), Reward: 0
[[ 0.  0.  0.  2.]
 [ 0.  2. -4.  0.]
 [ 0. -4.  0.  0.]
 [ 0.  6. -4.  9.]]
-----------------------------------------------------
Step No: 2
Current State: (1, 0), Action: 2 (Right), Reward: 0
[[ 0.  0.  0.  2.]
 [ 0.  2. -4.  0.]
 [ 0. -4.  0.  0.]
 [ 0.  6. -4.  9.]]
-----------------------------------------------------
Step No: 3
Current State: (2, 0), Action: 1 (Down), Reward: 0
[[ 0.  0.  0.  2.]
 [ 0.  2. -4.  0.]
 [ 0. -4.  0.  0.]
 [ 0.  6. -4.  9.]]
-----------------------------------------------------
Step No: 4
Current State: (2, 0), Action: 2 (Right), Reward: 0
[[ 0.  0.  0.  2.]
 [ 0.  2. -4.  0.]
 [ 0. -4.  0.  0.]
 [ 0.  6. -4.  9.]]
```

Ans to 3:

The observation space of the environment precisely represents the available states. To practice safe decision making, it ensures that the agent's perception of the world corresponds to the actual environment. Illegal moves like getting out of the designated grid are restricted. We simply used min, max function of python to make sure that the agent is always within its legal area. This is how we ensured the safety of our environment.

Report for PART- 2

Ans to ques 1:

SARSA:

Update Function: $Q(s, a) <- Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$

Key Features: In SARSA, Q-values are updated based on its current policy, which, in certain situations, makes it more stable. However, it may decrease the learning speed because it is influenced by its own exploratory behavior.

Advantages: It's stable and suitable for environments with high exploration.

Disadvantages: Converge slowly at times. Additionally, policy could be suboptimal due to its own exploratory actions.

Ans to ques 2:

Base case:

Number of episodes: 100

Epsilon decay: 0.96

I ran the environment for 100 episodes for non-greedy and 10 episodes for greedy and found out the total rewards. Chose the epsilon decay by the below formula.

decay factor = (Final epsilon value / Initial epsilon value) ^ (1/number of episodes)

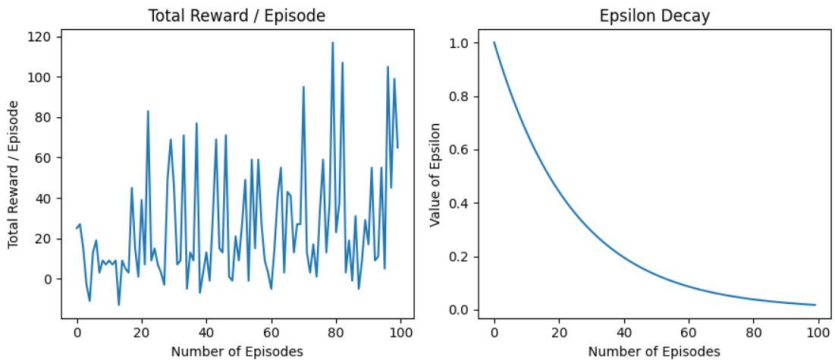
= (0.01 / 1) ^ (1/100) = 0.954 (I used 0.96 directly)

After running the non-greedy I found out that most of the time the player is achieving 0-60 points until reaching the final destination with +9 which is decent. Then, from the greedy one I found that it's learning to maximize the points with time. Moreover, it's filling out the Q-table adequately.

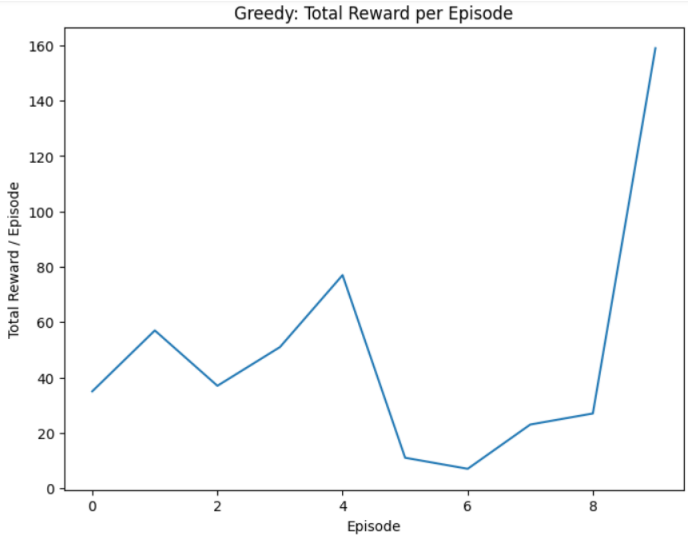
The initial and final Q-tables are included below.

```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[ 5.27920397  5.38294328 19.68958197  5.20626838]
 [ 0.64759092 17.00209183  2.4074954   5.78064481]
 [ 3.53234362  2.79443394 19.94707971  1.05864087]
 [22.12889598  5.50655739  5.35246158  1.00698659]
 [ 6.4416536   4.52085233 21.01999155  1.1911706 ]
 [ 1.03874103  2.83473973 18.3720079   2.55118115]
 [ 2.83447246  3.3846907  19.27470052  1.85451258]
 [ 2.75342049  2.61124147  5.4548898  16.94965392]
 [ 5.47623696 18.2928771   0.97050288  1.40497473]
 [ 0.44936735  0.9386692  21.76954708  3.55776115]
 [ 4.35502263 18.70584027  5.59155118  5.26548871]
 [ 4.83874776  0.02368808 17.03999285  4.19248782]
 [ 3.03768107 16.65615118 -0.3932704   3.33400969]
 [19.23032239  5.31916946  1.16490328  0.03164969]
 [17.92419376  7.38206144  2.9716388   3.00057523]
 [ 5.52921698 -0.65989649 21.40536124  8.22496879]]
```

Total Reward per episode and Epsilon Decay visualizations are given below.



Ran my environment for 10 episodes, where the agent chooses only greedy actions from the learned policy. The plot is given below where the total reward per episode is depicted.



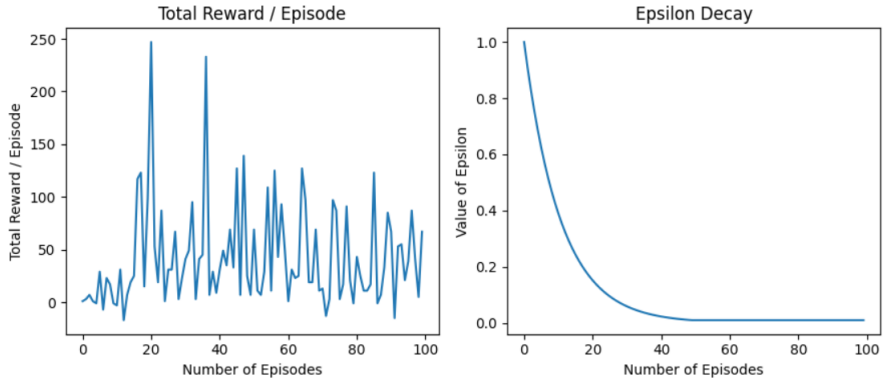
Ans to ques 3:

Hyperparameter Tuning - Case 1:

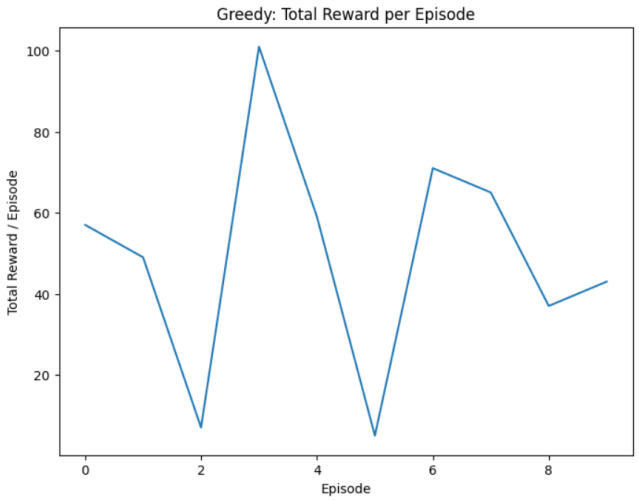
Number of episodes: 100

Epsilon decay: 0.91

In this case I decreased the epsilon decay from 0.96 to 0.91. The lower epsilon decay value has resulted in a slower decay. As a result, the agent/player explored more for a more extended period and gained more rewards on the go, as we see in the plots.




```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[ 7.37256799 34.61350383 -1.3355898   2.60057666]
 [ 8.3363749   2.85307267  1.61458468 31.65006042]
 [33.44832128  0.04859923  1.62534004  2.71685848]
 [ 4.47876902  1.41113014 34.37585071  2.29855427]
 [ 3.27957906  7.7877343  32.28616984  3.57268437]
 [ 7.58350084  2.69847256 32.23996588 -0.48608591]
 [-0.5109997   0.0792     34.9886018   3.75014178]
 [-0.82582924  1.62006608 33.50274778  3.38297038]
 [-0.99282025  5.868143   36.50628433 -0.30928571]
 [ 3.55339234 33.28637157  2.05397059 -0.34734648]
 [ 7.6530452   5.28693807 36.28976454 -0.40633809]
 [ 2.21669591 32.7455662   2.74157426  5.88434657]
 [ 3.36303244  8.19883508 32.73928974  8.79802491]
 [30.80153282  4.68148295  8.59494876 10.00584967]
 [ 2.73819283 32.5030276   9.86965921  0.80392859]
 [31.92003945  7.41466915  0.09543433  1.75794584]]
```

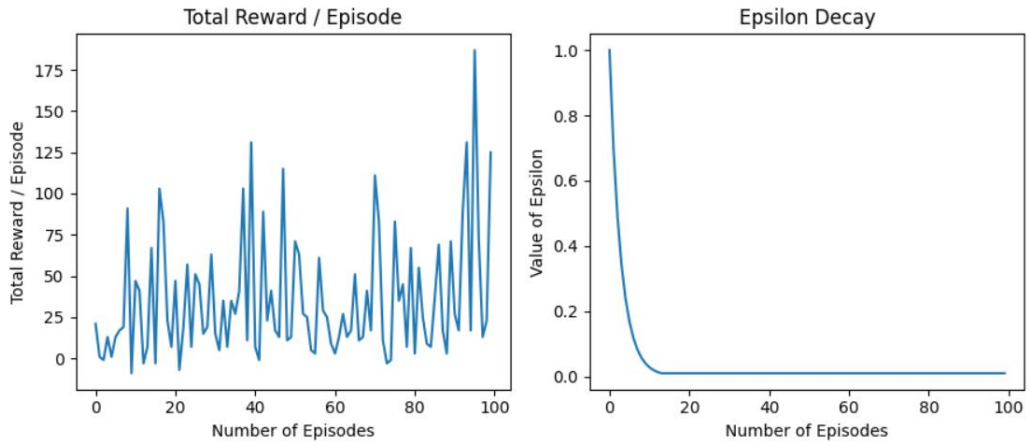


Hyperparameter Tuning - Case 2:

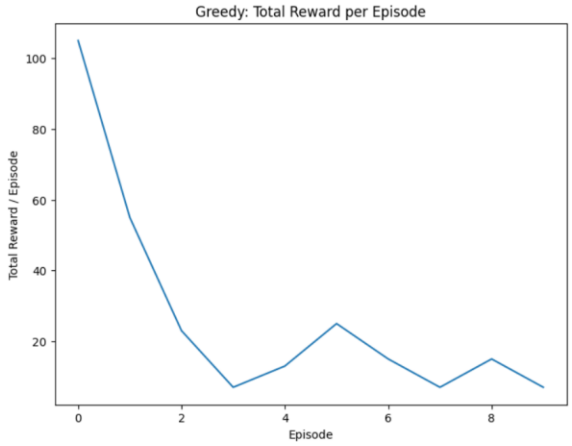
Number of episodes: 100

Epsilon decay: 0.70

In this case I decreased the epsilon decay even more from 0.96 to 0.7, which is really low. The lower epsilon decay value has resulted in a slower decay. As a result, the agent/player explored more for a more extended period and gained more rewards on the go, as we see in the plots.



```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[ 8.17583883e+00 -8.00000000e-01  2.82669202e+01 -6.69155045e-02]
 [ 2.79838516e+01  4.93369613e+00  5.06272657e-02  6.13563282e+00]
 [ 5.62539744e+00  2.41398861e+01  2.22610629e+00 -5.65565039e-01]
 [ 7.33452266e+00  1.74691180e-01  9.52135636e+00  2.91418380e+01]
 [ 7.24358289e+00 -6.13728276e-02  5.80480645e-01  2.95451488e+01]
 [ 1.46606217e+00 -9.83650314e-04  3.20791810e+00  2.87823940e+01]
 [-7.79504726e-01  2.16459601e+00  0.00000000e+00  2.85132009e+01]
 [ 7.85958574e-02  4.62805039e-01  2.62607742e+01 -3.13632000e-02]
 [ 3.09475963e+00  8.88201954e+00  2.95199809e+01  5.24970663e+00]
 [ 2.86230651e+01 -5.95520188e-01 -7.39267313e-01  4.86712321e+00]
 [ 5.24423602e+00  5.08295512e-01  5.03656324e+00  2.79587046e+01]
 [ 2.94223849e+01  4.73629895e-01  1.26953730e+00  8.02897920e-03]
 [ 2.80915861e+01  4.87609027e-01  6.23793311e+00  6.74758830e+00]
 [ 2.80672547e+01 -5.88250490e-01 -4.16324419e-01 -6.12470987e-01]
 [ 2.77888654e+01  4.96817836e+00  1.51808389e+00  4.99538377e+00]
 [ 2.75139622e+01  3.96254235e+00 -2.55099923e-01  1.05739972e+01]]
```

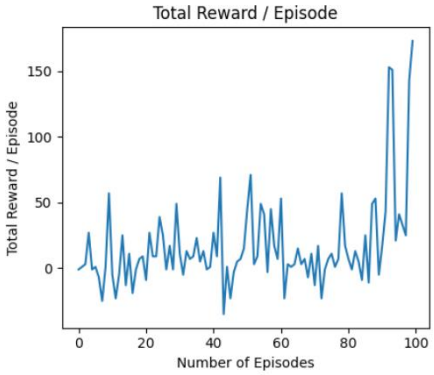


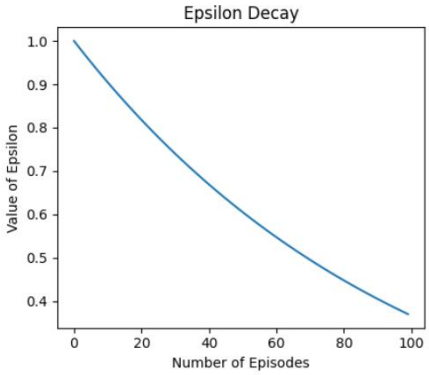
Hyperparameter Tuning - Case 3:

Number of episodes: 100

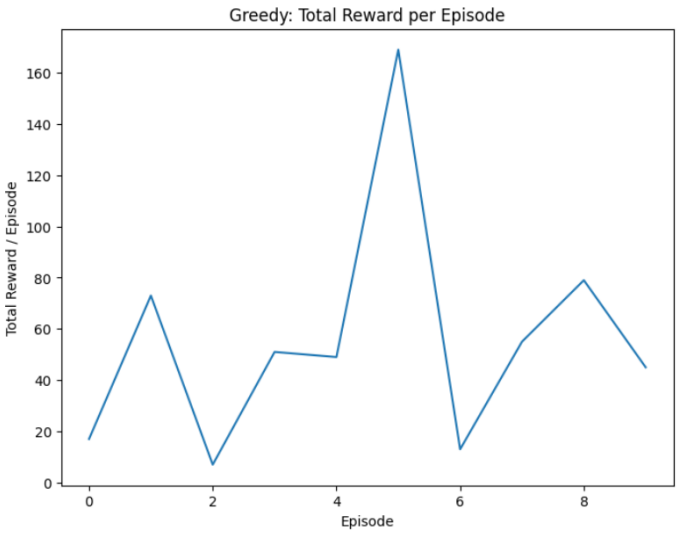
Epsilon decay: 0.99

In this case I increased the epsilon decay from 0.96 to 0.99, which is really low. The higher epsilon decay value has resulted in a faster decay. As a result, the agent/player explored less than the previous ones, as we see in the plots below.






```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[6.52053011 3.242086   2.93860076 5.01260095]
 [7.21832056 4.26935016 3.3960592  4.25251398]
 [7.5531493  4.73483605 3.20317593 4.9842384 ]
 [6.58010693 3.87002317 4.10658633 5.74362927]
 [6.38559007 2.20657661 5.20422043 4.07423218]
 [6.22192568 3.53891662 6.42093172 3.01655569]
 [5.94286118 3.23482356 7.52833132 7.14232085]
 [5.02044307 2.92327126 2.56193116 7.59888783]
 [7.30801885 3.89117227 3.16735346 5.32625573]
 [4.22349501 4.81415295 4.23212947 7.90414617]
 [6.49548467 2.33473679 4.92891069 4.52477013]
 [6.81557535 2.82949745 5.80880495 4.49630556]
 [8.46779499 3.83825328 4.01760054 6.90583118]
 [6.67826181 3.07635981 3.63070712 6.48857248]
 [4.5686235  3.83888686 4.79381645 5.89148331]
 [7.34455591 4.73585687 4.88253174 4.92400086]]
```

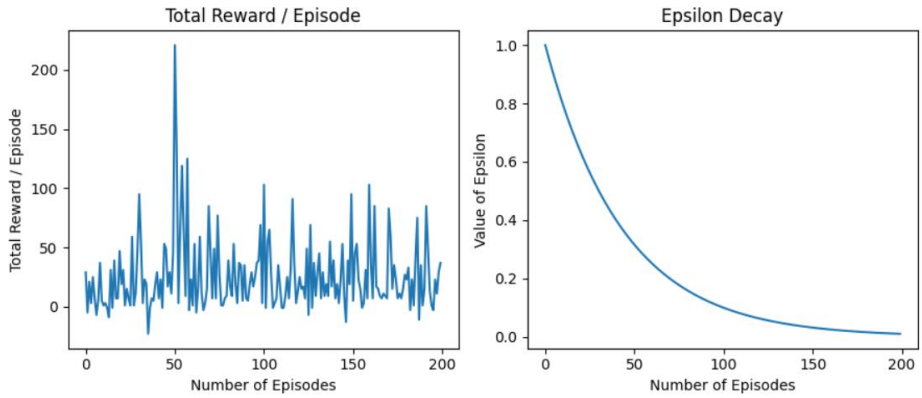


Hyperparameter Tuning - Case 4:

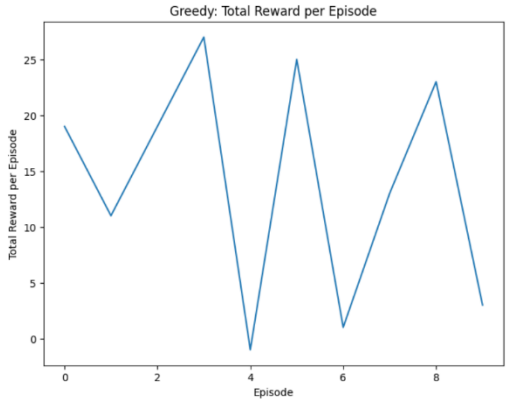
Number of episodes: 200

Epsilon decay: 0.9772

In this case I increased the number of episodes. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt more.




```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[31.71113381  8.73346312  6.76251288 12.35387102]
 [ 9.95768769 11.44782461 28.15689307  4.37441253]
 [13.53361218  6.87190938 29.53730212  6.89981687]
 [10.91236291  5.58518398 28.31858035  6.72142538]
 [11.62657223 28.69458139  9.26273389  4.69508338]
 [ 8.63219546 13.46549883 27.70748495 10.07341532]
 [11.56452095  7.89157292 26.91964386  6.02588651]
 [ 9.52738194 29.31692393  7.35130723 10.03695759]
 [33.02918517  9.44942297  6.6479534   9.21358147]
 [28.18245294 10.64961577 10.54012789 12.22230418]
 [ 5.43889551 29.08485141  7.27163974  6.37224953]
 [ 9.00284727 27.80952787  5.20700361 11.3999116 ]
 [14.19407505  7.75417535 28.51111099 10.96132244]
 [30.01419563  7.19628373  5.8774449  10.95380052]
 [31.58299078 12.8624289  16.01326143 10.91377949]
 [ 7.01114438 27.35988629  6.03249882  8.93083756]]
```

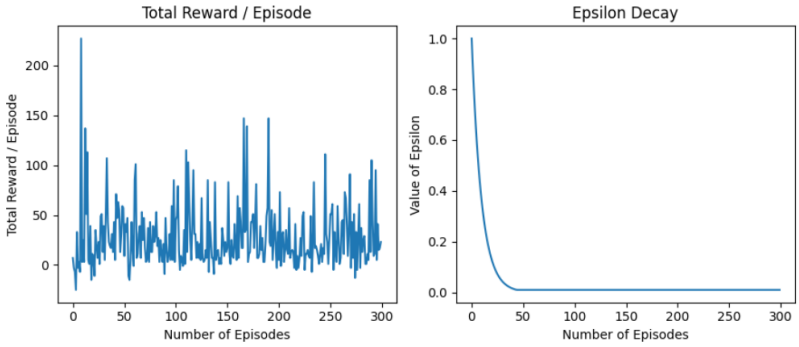


Hyperparameter Tuning - Case 5:

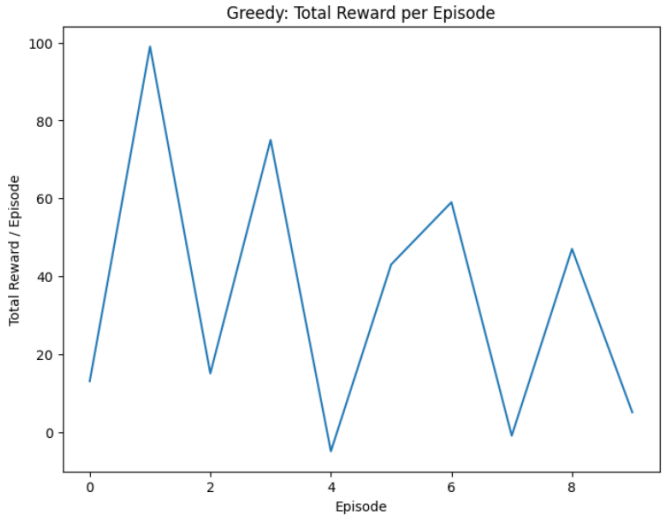
Number of episodes: 300

Epsilon decay: 0.90

In this case I increased the number of episodes even more. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt more.



```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[11.13197019 41.46863855  8.33569324  8.54427089]
 [19.04642422 43.28176645 18.97287442 20.05308039]
 [ 0.18359803 10.24487154 42.30708217 21.21939596]
 [ 0.16634225  8.35229355 41.74360699  4.28616457]
 [12.41534367  5.92819824 42.27766859 14.95309598]
 [17.51743677 15.8401733  42.08924508  8.97741308]
 [10.72324713  9.52478616 42.36691354  1.98494531]
 [42.17421811 19.43158756  6.36226565  4.91539041]
 [12.35986216  0.92563414  0.7709908  41.55318893]
 [44.77953975  7.48408661  0.70571126  1.75531907]
 [18.3928126  41.23790764 10.73546896  8.03636648]
 [16.83855714 42.87352373  8.9759126   0.6465914 ]
 [12.56477404  2.40708699 42.77867406 15.67241115]
 [45.37139235  6.56717506 13.67235749 -0.16797617]
 [ 7.66149017 41.78277843  1.87570201  7.98926038]
 [46.17820119 13.45873432  2.49809196  8.07589241]]
```

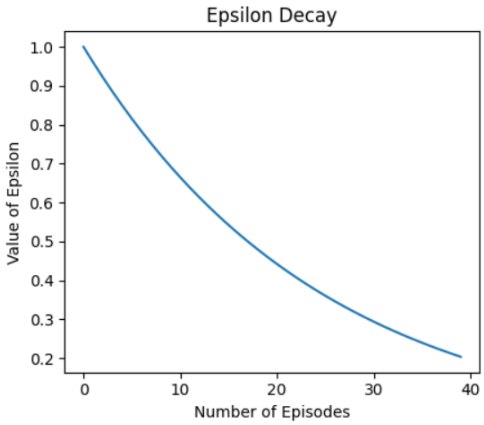


Hyperparameter Tuning - Case 6:

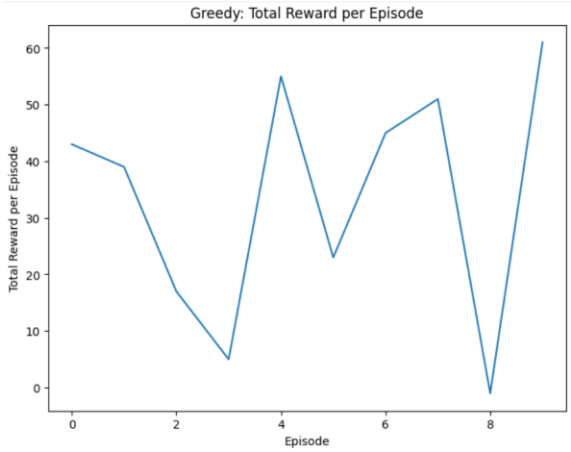
Number of episodes: 40

Epsilon decay: 0.96

In this case I decreased the number of episodes. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt less.




```
Q-table (At the beginning):
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]

Q-table (When trained):
[[ 0.32431204  0.96750558  5.13311793  0.02804539]
 [-0.91317283  0.75078227  3.89143685  0.34996394]
 [ 6.85932789  0.84787424 -0.0761609   1.50774247]
 [ 0.84890867  1.86992546  4.6599898   1.62916877]
 [ 3.11003297  1.21716668 -0.06339977 -0.61904287]
 [ 1.21960479  1.41817566  4.23458933  1.15026049]
 [ 5.80360571  0.18033199  1.25109717  0.23488892]
 [ 0.52898772  3.39420487  0.04643567  1.03688485]
 [ 3.12359641  0.61382678  3.97677512  0.90786381]
 [-0.16546808  2.09211466  1.89645299 -0.08864003]
 [ 6.16914882  0.99697475  2.34166763  1.1196509 ]
 [-0.8968097   0.99548583  4.25723193  0.96574817]
 [ 0.56126563  2.64476853  4.75308929  1.08976617]
 [ 1.50488908  0.99368138  3.39213496  0.64956379]
 [-0.40528207  2.86058889  2.80944163 -0.52330086]
 [ 0.02592483  3.11517102  0.56279168  0.30466562]]
```



Report for PART- 3

Ans to ques 1:

Double Q-learning:

Update Function: Double Q-learning updates one with the estimates from the other and alternate between two Q-tables. Moreover, it aids in reducing overestimation bias.

Key Features: Generates more accurate estimates by using two Q-tables to prevent overestimation of Q-values.

Advantage: Overestimation bias is addressed.

Disadvantage: Compared to implementation of SARSA, Double Q-learning is a bit more complicated.

Ans to ques 2:

Base case:

Number of episodes: 100

Epsilon decay: 0.96

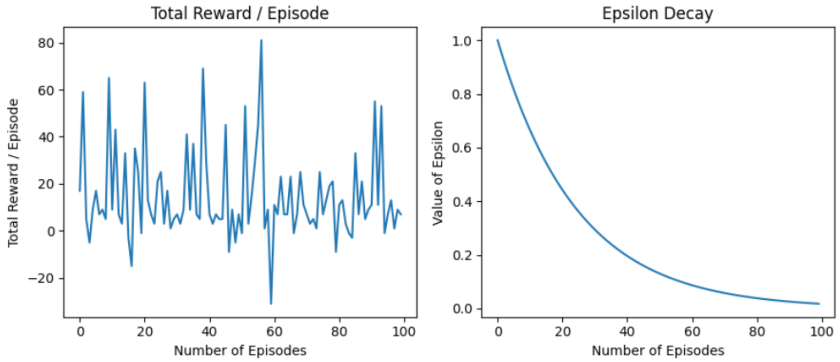
I ran the environment for 100 episodes for non-greedy and 10 episodes for greedy and found out the total rewards. Chose the epsilon decay by the below formula.

decay factor = (Final epsilon value / Initial epsilon value) ^ (1/number of episodes)

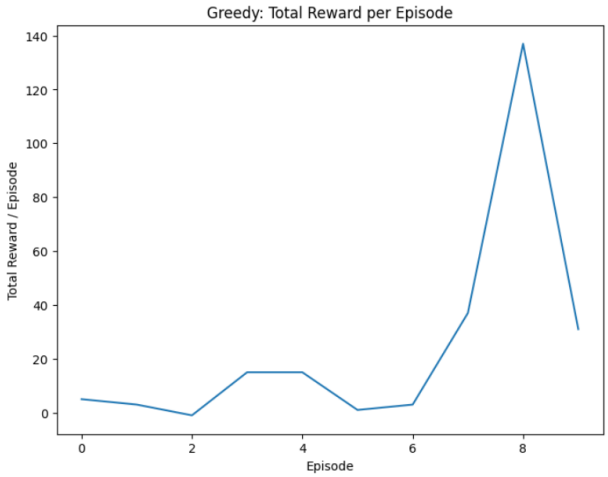
= (0.01 / 1) ^ (1/100) = 0.954 (I used 0.96 directly)

After running the non-greedy I found out that most of the time the player is achieving 0-50 points until reaching the destination with +9 which is decent. Then, from the greedy one I found out that it's learning to maximize the points with time. Moreover, it's filling out the Q-table adequately.

Total Reward per episode and Epsilon Decay visualizations are given below.



Ran my environment for 10 episodes, where the agent chooses only greedy actions from the learned policy. The plot is given below where the total reward per episode is depicted.



The initial and trained Q-tables are included below.

```
Q-table_A: At the beginning
[[ 3.25047547 -0.04829744  9.73905311 -0.27697564]
 [ 0.63366168  1.32282026  6.75547938  1.00744253]
 [ 5.76149372  0.70057792 -0.21346282 -1.21708675]
 [ 4.70878551  0.82741253  8.64116273  0.12956145]
 [ 8.8415471   1.65018963  1.71065277  1.60311772]
 [ 1.50278803  0.58814607  5.65747343  0.22369919]
 [ 1.37009926  7.25790014  1.07362248  1.68512949]
 [11.69976926  2.26221911  1.21316504  0.        ]
 [ 8.33297802  1.90067297  3.12772411 -0.23778855]
 [ 0.38027795  7.36904155  1.16087189 -0.60738986]
 [ 7.51502351  0.48333874 -0.08892875 -0.77303287]
 [-0.50207871  8.00716577  2.46139737 -0.23386392]
 [-0.18023859  1.89093047  7.5408287   0.04555154]
 [ 7.50784326  3.23908948 -0.60916421  0.59577615]
 [-0.84500015  0.40880916  9.69365664  0.88265455]
 [ 1.07750545  0.9788236   7.08150837  1.06648928]]

Q-table_B: At the beginning
[[-4.73747190e-02  7.33522978e-01  7.59571094e+00 -2.20377414e-03]
 [-7.09947801e-01  1.73798859e-01  7.04049438e+00  1.21511156e-01]
 [ 8.41846726e+00 -2.20798238e-01  5.18446889e-01  6.79778219e-01]
 [ 1.89194000e-01  1.51622216e+00  6.70613363e+00  1.00470557e+00]
 [ 6.03712823e+00  0.00000000e+00  3.90413885e+00  3.71998814e-01]
 [ 7.51444617e-01  1.67867166e-01  7.04027041e+00  2.71077485e+00]
 [ 1.11900351e-01  6.39274242e+00  1.06437381e+00  5.12399643e-01]
 [ 6.71038053e+00  1.93250957e+00  1.64363633e+00  5.27008409e-01]
 [ 1.04607042e+01 -1.24863356e+00  9.36997620e-01  8.56595759e-01]
 [-8.57442807e-01  6.83798632e+00  2.67973035e+00  1.23848806e+00]
 [ 7.27444780e+00  1.77199467e+00 -3.39815373e-01  8.78082989e-01]
 [ 3.70678006e+00  1.01595537e+01  1.09846097e+00 -1.25072288e+00]
 [ 2.00853957e-01  8.54960161e-01  7.61525401e+00  8.72473726e-01]
 [ 6.41847536e+00 -1.64850943e-01  1.72408466e+00 -1.56635838e+00]
 [ 3.25577092e+00  2.08008075e+00  7.41157065e+00  1.78674984e+00]
 [ 2.82115310e+00  3.69213662e+00  8.29151047e+00  1.54840285e+00]]
```

```
Trained Q-table_A:
[[ 3.25047547 -0.04829744  9.73905311 -0.27697564]
 [ 0.63366168  1.32282026  6.75547938  1.00744253]
 [ 5.76149372  0.70057792 -0.21346282 -1.21708675]
 [ 4.70878551  0.82741253  8.64116273  0.12956145]
 [ 8.8415471   1.65018963  1.71065277  1.60311772]
 [ 1.50278803  0.58814607  5.65747343  0.22369919]
 [ 1.37009926  7.25790014  1.07362248  1.68512949]
 [11.69976926  2.26221911  1.21316504  0.        ]
 [ 8.33297802  1.90067297  3.12772411 -0.23778855]
 [ 0.38027795  7.36904155  1.16087189 -0.60738986]
 [ 7.51502351  0.48333874 -0.08892875 -0.77303287]
 [-0.50207871  8.00716577  2.46139737 -0.23386392]
 [-0.18023859  1.89093047  7.5408287   0.04555154]
 [ 7.50784326  3.23908948 -0.60916421  0.59577615]
 [-0.84500015  0.40880916  9.69365664  0.88265455]
 [ 1.07750545  0.9788236   7.08150837  1.06648928]]

Trained Q-table_B:
[[-4.73747190e-02  7.33522978e-01  7.59571094e+00 -2.20377414e-03]
 [-7.09947801e-01  1.73798859e-01  7.04049438e+00  1.21511156e-01]
 [ 8.41846726e+00 -2.20798238e-01  5.18446889e-01  6.79778219e-01]
 [ 1.89194000e-01  1.51622216e+00  6.70613363e+00  1.00470557e+00]
 [ 6.03712823e+00  0.00000000e+00  3.90413885e+00  3.71998814e-01]
 [ 7.51444617e-01  1.67867166e-01  7.04027041e+00  2.71077485e+00]
 [ 1.11900351e-01  6.39274242e+00  1.06437381e+00  5.12399643e-01]
 [ 6.71038053e+00  1.93250957e+00  1.64363633e+00  5.27008409e-01]
 [ 1.04607042e+01 -1.24863356e+00  9.36997620e-01  8.56595759e-01]
 [-8.57442807e-01  6.83798632e+00  2.67973035e+00  1.23848806e+00]
 [ 7.27444780e+00  1.77199467e+00 -3.39815373e-01  8.78082989e-01]
 [ 3.70678006e+00  1.01595537e+01  1.09846097e+00 -1.25072288e+00]
 [ 2.00853957e-01  8.54960161e-01  7.61525401e+00  8.72473726e-01]
 [ 6.41847536e+00 -1.64850943e-01  1.72408466e+00 -1.56635838e+00]
 [ 3.25577092e+00  2.08008075e+00  7.41157065e+00  1.78674984e+00]
 [ 2.82115310e+00  3.69213662e+00  8.29151047e+00  1.54840285e+00]]
```

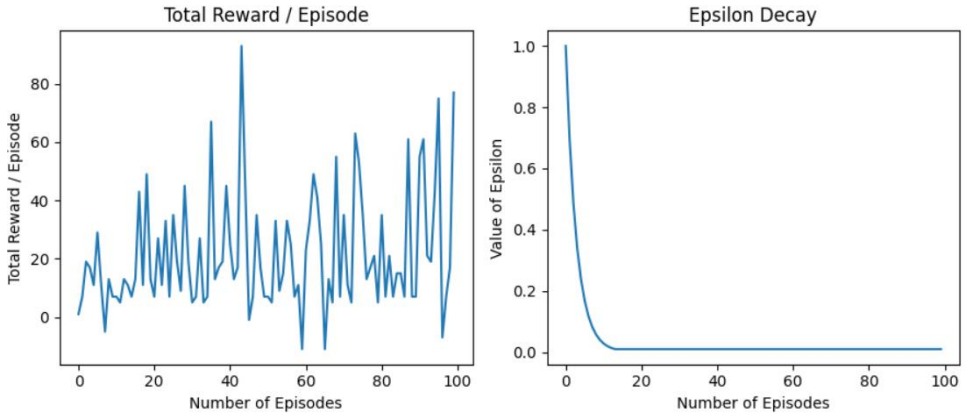
Ans to ques 3:

Hyperparameter Tuning - Case 1:

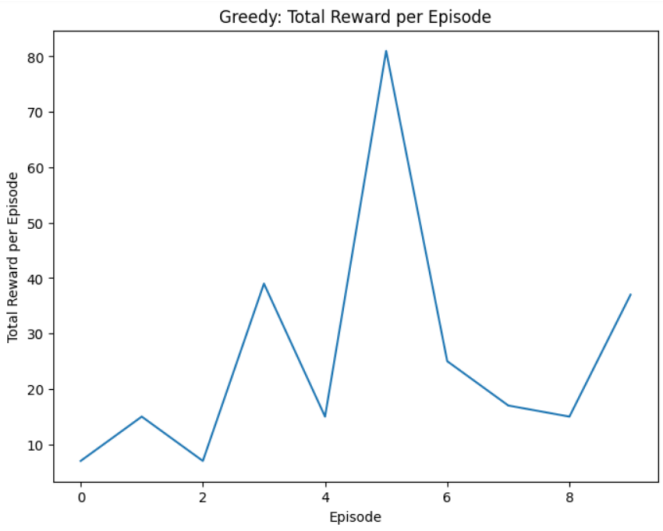
Number of episodes: 100

Epsilon decay: 0.70

In this case I decreased the epsilon decay so much from 0.96 to 0.7, which is really low. The lower epsilon decay value has resulted in a slower decay. As a result, the agent/player explored more for a more extended period and gained more rewards on the go, as we see in the plots.


Total Reward / Episode
Total Reward / Episode
Number of Episodes
0
20
40
60
80
100

Epsilon Decay
Value of Epsilon
Number of Episodes
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


Greedy: Total Reward per Episode
Total Reward per Episode
Episode
10
20
30
40
50
60
70
80
0
2
4
6
8

```
Q-table_A: At the beginning
[[12.28315431  1.54841889  2.84844744  0.        ]
 [12.3593447   0.          1.67935834  0.        ]
 [ 1.1104325   0.         12.59717496  0.        ]
 [12.43314831  0.         -0.34648251  0.        ]
 [-0.749312   -0.63477802  1.386469   12.80251449]
 [ 0.61499341  0.         11.27877821  2.60887761]
 [12.38623469 -0.05505775 -1.44        0.        ]
 [11.84366152  0.04982397  0.          0.03991217]
 [ 0.          0.27340971  0.         12.76115142]
 [ 0.         -0.8        13.07662463  0.27834829]
 [-0.12672     0.         12.15719904  1.12907073]
 [-0.5313536   1.59162724 -0.12165171 12.28972994]
 [12.76163674  0.          0.          0.        ]
 [12.72253523  0.          0.          0.        ]
 [12.50690751  0.1426187   0.          0.        ]
 [12.14612383  0.          0.          0.        ]]

Q-table_B: At the beginning
[[11.96970559  0.          0.          0.        ]
 [12.15547777  0.          0.          0.        ]
 [ 0.          1.88630295 11.29696088  0.        ]
 [11.85065511 -0.10520801  0.          0.        ]
 [ 0.          0.          0.58577749 12.1218856 ]
 [ 0.          0.         12.0486099   0.17731926]
 [11.11336167  0.          0.         -0.61440256]
 [13.67006996  0.          0.          0.        ]
 [-0.02509056 -0.64       -0.64       12.77266264]
 [-0.48        0.         14.28350655  1.63863589]
 [ 0.         -0.8        13.50762612  0.        ]
 [ 0.4         1.39929902  0.         12.59546061]
 [11.72426841  0.          0.          0.        ]
 [11.2616065   0.         -0.8         0.        ]
 [11.71393103  0.          0.          0.        ]
 [12.91136602  0.          0.          0.        ]]
```

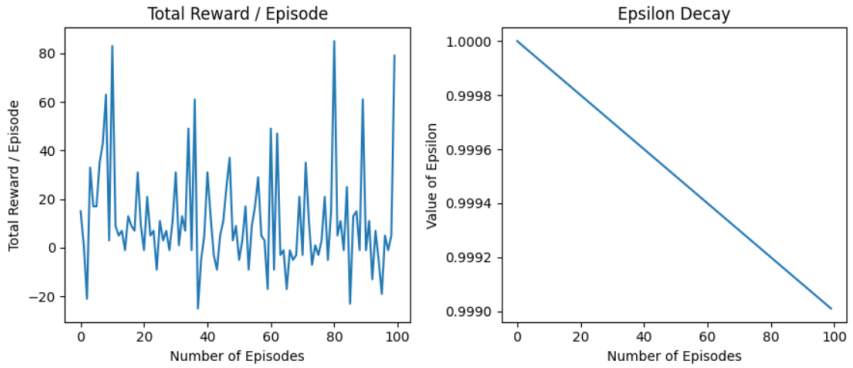
```
Trained Q-table_A:
[[12.28315431  1.54841889  2.84844744  0.        ]
 [12.3593447   0.          1.67935834  0.        ]
 [ 1.1104325   0.         12.59717496  0.        ]
 [12.43314831  0.         -0.34648251  0.        ]
 [-0.749312   -0.63477802  1.386469   12.80251449]
 [ 0.61499341  0.         11.27877821  2.60887761]
 [12.38623469 -0.05505775 -1.44        0.        ]
 [11.84366152  0.04982397  0.          0.03991217]
 [ 0.          0.27340971  0.         12.76115142]
 [ 0.         -0.8        13.07662463  0.27834829]
 [-0.12672     0.         12.15719904  1.12907073]
 [-0.5313536   1.59162724 -0.12165171 12.28972994]
 [12.76163674  0.          0.          0.        ]
 [12.72253523  0.          0.          0.        ]
 [12.50690751  0.1426187   0.          0.        ]
 [12.14612383  0.          0.          0.        ]]

Trained Q-table_B:
[[11.96970559  0.          0.          0.        ]
 [12.15547777  0.          0.          0.        ]
 [ 0.          1.88630295 11.29696088  0.        ]
 [11.85065511 -0.10520801  0.          0.        ]
 [ 0.          0.          0.58577749 12.1218856 ]
 [ 0.          0.         12.0486099   0.17731926]
 [11.11336167  0.          0.         -0.61440256]
 [13.67006996  0.          0.          0.        ]
 [-0.02509056 -0.64       -0.64       12.77266264]
 [-0.48        0.         14.28350655  1.63863589]
 [ 0.         -0.8        13.50762612  0.        ]
```

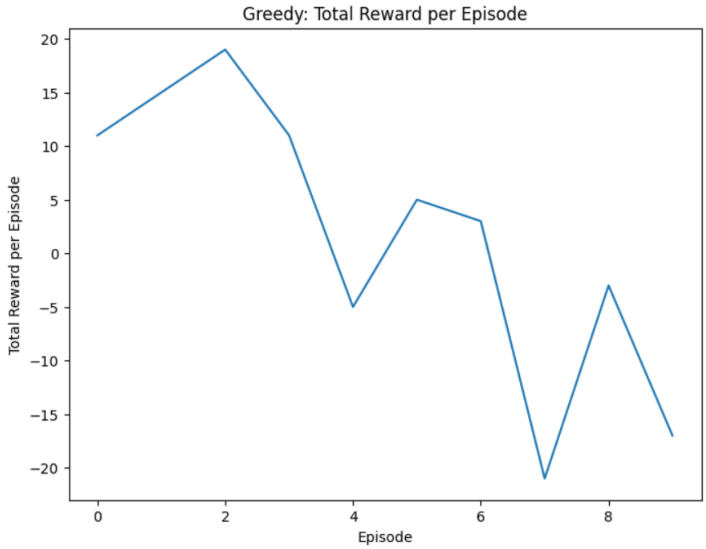
<u>Hyperparameter Tuning - Case 2:</u>

Number of episodes: 100

Epsilon decay: 0.99999

In this case I increased the epsilon decay from 0.96 to 0.99999, which is really low. The higher epsilon decay value has resulted in a faster decay. As a result, the agent/player explored less than the previous ones, as we see in the plots below.






```
Q-table_A: At the beginning
[[4.36554651 3.80016425 6.61271212 3.4630905 ]
 [1.86535797 3.05478033 4.1837618  3.52111713]
 [5.60596221 4.70406323 4.53652085 2.26772096]
 [3.84771585 5.15751032 4.22381302 3.75173294]
 [2.65645325 3.79413526 3.34199086 2.08039263]
 [5.25356206 3.53541881 4.1936082  3.01003366]
 [3.53513312 3.68677558 2.76411391 3.01795592]
 [2.90089    3.52147336 2.75434179 2.31067446]
 [3.62845656 3.97987182 4.18534096 4.30489956]
 [2.96928239 2.7024201  3.20716718 3.6245496 ]
 [2.67393246 3.57715864 2.18695528 3.67929486]
 [2.79277973 2.67665175 2.37757248 4.56640078]
 [2.90000383 3.29870959 4.16608544 3.5768837 ]
 [2.9583578  3.69294003 4.39385437 2.87287525]
 [2.92444159 3.85114952 2.66377308 3.63656111]
 [4.7720696  3.97882419 4.40329114 3.94266104]]

Q-table_B: At the beginning
[[2.76583131 2.93797262 3.75407373 3.19470045]
 [3.13129959 3.47016592 5.39714582 3.36092247]
 [3.34868204 2.75249335 2.9168913  2.77089766]
 [3.84549758 4.30766335 3.91851824 4.04283125]
 [4.66787847 3.25640445 4.57829025 3.09118406]
 [2.4387451  2.83387564 3.68744183 3.44842854]
 [5.52113435 2.33075275 5.11402669 3.85955768]
 [6.04476444 3.93355302 4.68148318 3.01067061]
 [3.06928249 3.12946245 4.75393713 3.09100214]
 [3.44455626 3.41831404 3.37690132 3.31637697]
 [3.30368308 2.30216574 1.54766326 3.21454046]
 [4.42316487 4.5075494  3.10929665 4.27584285]
 [3.88488413 4.25773952 5.21039339 3.95979232]
 [4.63214111 4.61802702 4.63142225 2.29116198]
 [5.03610923 2.58674755 4.87186544 3.65485927]
 [4.82172377 3.0397316  1.0244128  2.47917657]]
```

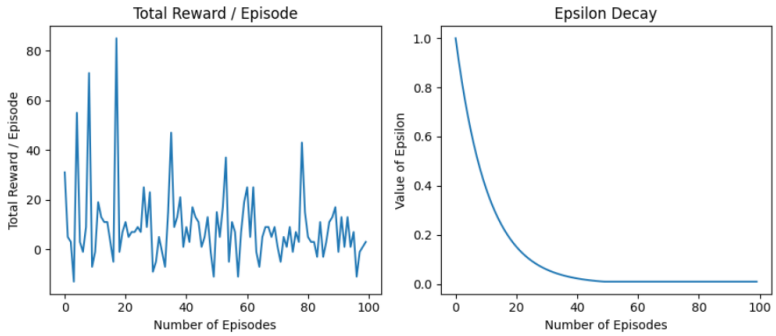
```
Trained Q-table_A:
[[4.36554651 3.80016425 6.61271212 3.4630905 ]
 [1.86535797 3.05478033 4.1837618  3.52111713]
 [5.60596221 4.70406323 4.53652085 2.26772096]
 [3.84771585 5.15751032 4.22381302 3.75173294]
 [2.65645325 3.79413526 3.34199086 2.08039263]
 [5.25356206 3.53541881 4.1936082  3.01003366]
 [3.53513312 3.68677558 2.76411391 3.01795592]
 [2.90089    3.52147336 2.75434179 2.31067446]
 [3.62845656 3.97987182 4.18534096 4.30489956]
 [2.96928239 2.7024201  3.20716718 3.6245496 ]
 [2.67393246 3.57715864 2.18695528 3.67929486]
 [2.79277973 2.67665175 2.37757248 4.56640078]
 [2.90000383 3.29870959 4.16608544 3.5768837 ]
 [2.9583578  3.69294003 4.39385437 2.87287525]
 [2.92444159 3.85114952 2.66377308 3.63656111]
 [4.7720696  3.97882419 4.40329114 3.94266104]]

Trained Q-table_B:
[[2.76583131 2.93797262 3.75407373 3.19470045]
 [3.13129959 3.47016592 5.39714582 3.36092247]
 [3.34868204 2.75249335 2.9168913  2.77089766]
 [3.84549758 4.30766335 3.91851824 4.04283125]
 [4.66787847 3.25640445 4.57829025 3.09118406]
 [2.4387451  2.83387564 3.68744183 3.44842854]
 [5.52113435 2.33075275 5.11402669 3.85955768]
 [6.04476444 3.93355302 4.68148318 3.01067061]
 [3.06928249 3.12946245 4.75393713 3.09100214]
 [3.44455626 3.41831404 3.37690132 3.31637697]
 [3.30368308 2.30216574 1.54766326 3.21454046]
 [4.42316487 4.5075494  3.10929665 4.27584285]
 [3.88488413 4.25773952 5.21039339 3.95979232]
 [4.63214111 4.61802702 4.63142225 2.29116198]
 [5.03610923 2.58674755 4.87186544 3.65485927]
 [4.82172377 3.0397316  1.0244128  2.47917657]]
```

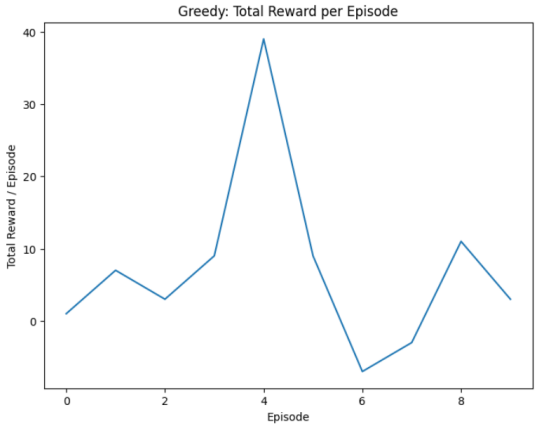
Hyperparameter Tuning - Case 3:

Number of episodes: 100

Epsilon decay: 0.91

In this case I decreased the epsilon decay from 0.96 to 0.91, which is really low. The lower epsilon decay value has resulted in a slower decay. As a result, the agent/player explored more for a more extended period and gained more rewards on the go, as we see in the plots.






```
Q-table_A: At the beginning
[[ 2.87354471  0.68362361 -0.1897632   0.552064  ]
 [ 4.52430552  1.2          0.97834193  0.        ]
 [-0.94914075  4.44085044 -0.674586   -0.32891106]
 [ 0.67963915  3.26061105 -0.69789389 -0.43773762]
 [ 3.60959798 -0.12672     0.3686368  -0.36672   ]
 [ 5.2675399   0.06346052 -0.75787085 -1.20850144]
 [ 5.08776358 -0.21485376  0.18505126  0.92953748]
 [-0.44491113 -0.20884465  2.37051186  0.78469199]
 [ 0.          0.          3.39107395 -1.04990087]
 [ 2.77510709  1.02836887  0.16306787 -0.1584    ]
 [ 0.          3.89413513  0.03440174  0.11260137]
 [-0.8313632   4.01368824 -0.43305993  0.        ]
 [ 5.14960913  0.29853214  0.38789223 -1.3277076 ]
 [ 3.74110267  0.         -0.30002842  0.22410987]
 [-0.45442333  3.74287013  0.35191959  0.        ]
 [-1.49171508  0.07596121  6.41774352  0.        ]]

Q-table_B: At the beginning
[[ 2.79852382e+00  2.95192417e-01 -3.52452063e-01  2.48396544e-03]
 [ 3.89271420e+00  1.62083558e+00  4.11150953e-01  0.00000000e+00]
 [ 0.00000000e+00  3.75029244e+00  1.75725476e-01 -1.21415936e+00]
 [-8.74776835e-01  3.39447729e+00  4.67863747e-01 -3.86618938e-03]
 [ 3.46510985e+00  5.52136455e-01  2.42348130e-01  6.71986360e-01]
 [ 4.51116986e+00  6.14701705e-01  2.86071963e-01  2.57239066e-02]
 [ 4.46624001e+00  3.82694400e-02 -4.92207804e-01 -5.97840447e-01]
 [ 2.05305776e+00 -4.70401914e-01  2.23480705e+00 -7.98263444e-01]
 [-4.60618868e-01 -2.02511163e-01  4.45008006e+00  0.00000000e+00]
 [ 4.09880652e+00 -2.92604631e-01  0.00000000e+00  0.00000000e+00]
 [-8.91251801e-01  4.45026724e+00  2.41066868e+00 -1.11213550e+00]
 [-1.61584481e+00  1.98566504e+00 -4.53607931e-01 -8.00000000e-01]
 [ 2.48273749e+00  0.00000000e+00 -1.54544566e+00  3.05419654e-01]
 [ 4.32086110e+00  0.00000000e+00  1.37224900e+00 -2.41564500e-01]
 [-8.25090560e-01  3.20054051e+00  1.50403206e-02  0.00000000e+00]
 [-3.45644876e-01  8.84688247e-01  4.27501659e+00 -4.53197255e-01]]
```

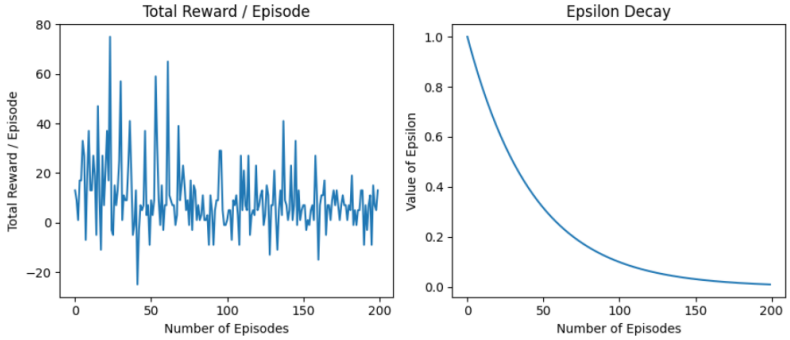
```
Trained Q-table_A:
[[ 2.87354471  0.68362361 -0.1897632   0.552064  ]
 [ 4.52430552  1.2          0.97834193  0.        ]
 [-0.94914075  4.44085044 -0.674586   -0.32891106]
 [ 0.67963915  3.26061105 -0.69789389 -0.43773762]
 [ 3.60959798 -0.12672     0.3686368  -0.36672   ]
 [ 5.2675399   0.06346052 -0.75787085 -1.20850144]
 [ 5.08776358 -0.21485376  0.18505126  0.92953748]
 [-0.44491113 -0.20884465  2.37051186  0.78469199]
 [ 0.          0.          3.39107395 -1.04990087]
 [ 2.77510709  1.02836887  0.16306787 -0.1584    ]
 [ 0.          3.89413513  0.03440174  0.11260137]
 [-0.8313632   4.01368824 -0.43305993  0.        ]
 [ 5.14960913  0.29853214  0.38789223 -1.3277076 ]
 [ 3.74110267  0.         -0.30002842  0.22410987]
 [-0.45442333  3.74287013  0.35191959  0.        ]
 [-1.49171508  0.07596121  6.41774352  0.        ]]

Trained Q-table_B:
[[ 2.79852382e+00  2.95192417e-01 -3.52452063e-01  2.48396544e-03]
 [ 3.89271420e+00  1.62083558e+00  4.11150953e-01  0.00000000e+00]
 [ 0.00000000e+00  3.75029244e+00  1.75725476e-01 -1.21415936e+00]
 [-8.74776835e-01  3.39447729e+00  4.67863747e-01 -3.86618938e-03]
 [ 3.46510985e+00  5.52136455e-01  2.42348130e-01  6.71986360e-01]
 [ 4.51116986e+00  6.14701705e-01  2.86071963e-01  2.57239066e-02]
 [ 4.46624001e+00  3.82694400e-02 -4.92207804e-01 -5.97840447e-01]
 [ 2.05305776e+00 -4.70401914e-01  2.23480705e+00 -7.98263444e-01]
 [-4.60618868e-01 -2.02511163e-01  4.45008006e+00  0.00000000e+00]
 [ 4.09880652e+00 -2.92604631e-01  0.00000000e+00  0.00000000e+00]
 [-8.91251801e-01  4.45026724e+00  2.41066868e+00 -1.11213550e+00]
 [-1.61584481e+00  1.98566504e+00 -4.53607931e-01 -8.00000000e-01]
 [ 2.48273749e+00  0.00000000e+00 -1.54544566e+00  3.05419654e-01]
 [ 4.32086110e+00  0.00000000e+00  1.37224900e+00 -2.41564500e-01]
 [-8.25090560e-01  3.20054051e+00  1.50403206e-02  0.00000000e+00]
 [-3.45644876e-01  8.84688247e-01  4.27501659e+00 -4.53197255e-01]]
```

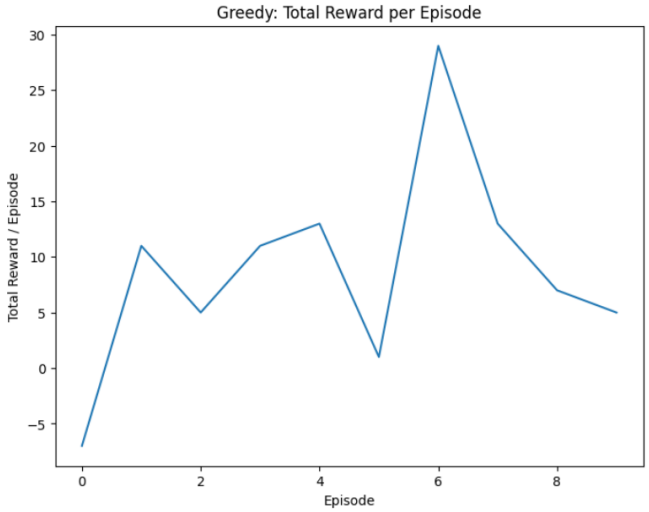
Hyperparameter Tuning - Case 4:

Number of episodes: 200

Epsilon decay: 0.9772

In this case I increased the number of episodes. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt more.





```
Q-table_A: At the beginning
[[ 2.46556145  1.42920635  0.80526705  2.62187682]
 [ 2.8556319   1.08350298  4.48842829  0.56766095]
 [ 4.67727066  1.87811925  2.68701756  0.75106385]
 [ 1.4155508   1.29898968  4.57705047  0.4340265 ]
 [ 4.37021807  1.16667821 -1.53605687  0.89619517]
 [ 2.80912738  0.29127929  6.32887769  1.65784268]
 [-0.68045808  0.10937338  7.15438516  0.43570224]
 [ 5.34135008  1.18634825  1.99642793  0.75496262]
 [ 3.58028975 -0.43532533  3.23433838  0.31620522]
 [ 2.51611328  1.0364588   0.20286503 -0.5263279 ]
 [ 4.03211994  2.66139302  2.33929146  1.10762524]
 [ 0.01472034 -0.85546052  3.57218816  1.13519601]
 [ 2.1188192   4.18457784  1.21889422  1.07539037]
 [ 1.65074594  1.51817134  4.36923558  0.10178767]
 [ 4.55549796  2.99400106  1.68947745  0.70875995]
 [ 4.20461114  2.11581661 -0.28406041  0.48954567]]

Q-table_B: At the beginning
[[ 0.63586771  2.5260815   0.2884143   3.72785224]
 [ 3.78072527 -1.09591285  7.37879951  1.72028091]
 [ 3.40321071  0.20367236 -0.97610597  0.26344185]
 [ 0.81928934  1.01522783  3.70045406 -0.48285093]
 [ 4.28693287  2.89867845  5.30006561  0.59900408]
 [ 0.06225763  2.21284289  6.64407849  1.8649946 ]
 [-1.3778498   0.51440093  2.92386881  1.22692629]
 [ 4.94765075  1.54700557  0.34858015  0.32854982]
 [ 3.95376895  2.08472474 -0.16982686  0.43576893]
 [ 4.97999825  1.19397317  1.28515613  0.36182125]
 [ 3.88843373  0.76994871  0.71273619  1.11603076]
 [ 0.71775855  1.3173138   5.81772329 -0.42419144]
 [ 1.26755923  1.24502422  1.07687231  0.1506105 ]
 [ 1.31390499  0.13837105  3.56002916 -0.72126179]
 [ 2.94618324  0.55183914  1.10481894  0.54179175]
 [ 6.4992358   2.07636789  1.23934805  1.42939879]]
```

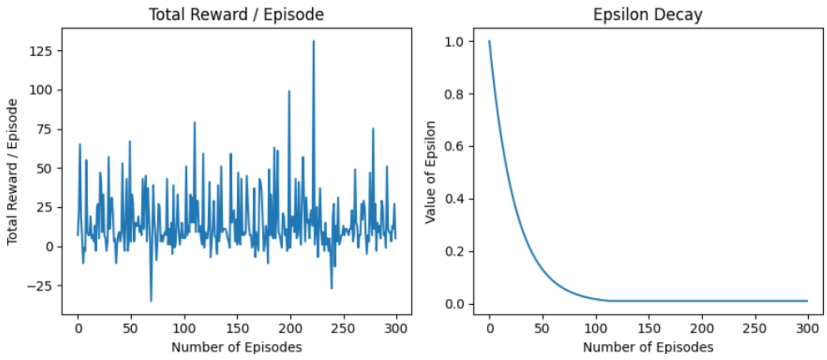
```
Trained Q-table_A:
[[ 2.46556145  1.42920635  0.80526705  2.62187682]
 [ 2.8556319   1.08350298  4.48842829  0.56766095]
 [ 4.67727066  1.87811925  2.68701756  0.75106385]
 [ 1.4155508   1.29898968  4.57705047  0.4340265 ]
 [ 4.37021807  1.16667821 -1.53605687  0.89619517]
 [ 2.80912738  0.29127929  6.32887769  1.65784268]
 [-0.68045808  0.10937338  7.15438516  0.43570224]
 [ 5.34135008  1.18634825  1.99642793  0.75496262]
 [ 3.58028975 -0.43532533  3.23433838  0.31620522]
 [ 2.51611328  1.0364588   0.20286503 -0.5263279 ]
 [ 4.03211994  2.66139302  2.33929146  1.10762524]
 [ 0.01472034 -0.85546052  3.57218816  1.13519601]
 [ 2.1188192   4.18457784  1.21889422  1.07539037]
 [ 1.65074594  1.51817134  4.36923558  0.10178767]
 [ 4.55549796  2.99400106  1.68947745  0.70875995]
 [ 4.20461114  2.11581661 -0.28406041  0.48954567]]

Trained Q-table_B:
[[ 0.63586771  2.5260815   0.2884143   3.72785224]
 [ 3.78072527 -1.09591285  7.37879951  1.72028091]
 [ 3.40321071  0.20367236 -0.97610597  0.26344185]
 [ 0.81928934  1.01522783  3.70045406 -0.48285093]
 [ 4.28693287  2.89867845  5.30006561  0.59900408]
 [ 0.06225763  2.21284289  6.64407849  1.8649946 ]
 [-1.3778498   0.51440093  2.92386881  1.22692629]
 [ 4.94765075  1.54700557  0.34858015  0.32854982]
 [ 3.95376895  2.08472474 -0.16982686  0.43576893]
 [ 4.97999825  1.19397317  1.28515613  0.36182125]
 [ 3.88843373  0.76994871  0.71273619  1.11603076]
 [ 0.71775855  1.3173138   5.81772329 -0.42419144]
 [ 1.26755923  1.24502422  1.07687231  0.1506105 ]
 [ 1.31390499  0.13837105  3.56002916 -0.72126179]
 [ 2.94618324  0.55183914  1.10481894  0.54179175]
 [ 6.4992358   2.07636789  1.23934805  1.42939879]]
```

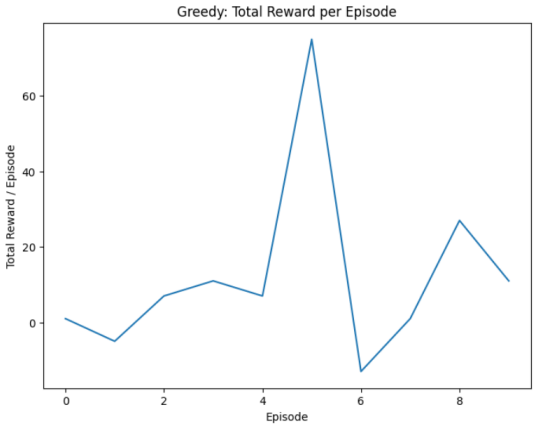
<u>Hyperparameter Tuning - Case 5:</u>

Number of episodes: 300

Epsilon decay: 0.96

In this case I increased the number of episodes. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt more.






```
Q-table_A: At the beginning
[[ 2.37403482e+01 -3.43253854e-01  4.46029039e-01  1.92070475e+00]
 [ 4.65384867e-01  7.36041192e+00  2.23697815e+01  2.85232175e-01]
 [ 1.81742353e+00 -8.61542185e-02  2.21963820e+01  2.02727599e-02]
 [ 2.22027823e+01  1.40392428e-01  6.75540232e-01 -3.83484844e-01]
 [ 1.41233053e+00  2.26937790e+01  3.99534457e-02  7.82866433e-01]
 [ 2.17972880e+01 -4.74753740e-01 -4.10861537e-01  7.08906680e-01]
 [-5.66337908e-01  7.42687670e-01  2.12329619e+01 -7.82993990e-02]
 [ 1.55507755e+00  1.76618927e+00  2.19127581e+01  7.52527356e-01]
 [-2.37509346e+00  2.37952168e+01 -5.59406351e-01  7.93430423e-01]
 [ 3.75624537e+00 -6.34597555e-01  2.50531852e+01  3.18521063e+00]
 [ 4.69221965e-01  2.25166494e+01  2.07114387e+00  6.38709775e-01]
 [ 2.26864463e+01  2.28787183e-01 -1.28691343e+00 -9.20537456e-01]
 [ 2.54666865e+01  2.85247994e-01  5.60030441e+00  2.36807459e+00]
 [ 1.70513950e+00 -3.81277170e-02  2.05136084e+01  8.92841489e-01]
 [ 3.85461958e-01  2.29426740e+01  2.69729793e+00 -7.23426594e-01]
 [ 2.33671658e+01 -2.20474945e-01  1.85525450e+00  2.84282000e+00]]

Q-table_B: At the beginning
[[ 2.46100794e+01  6.01217661e+00  7.38397274e-01  4.01725355e+00]
 [ 2.37231739e+00  3.15278850e-01  2.29476274e+01 -5.13881295e-01]
 [ 1.26374428e+00  8.23089434e-01  2.34073089e+01 -4.17153945e-01]
 [ 2.19418104e+01  1.82340132e+00  8.46033828e+00 -3.24926756e-01]
 [-2.97684612e+00  2.20912556e+01  5.28325404e+00  8.26539403e-01]
 [ 2.47734581e+01  2.07905339e+00  4.48766687e+00  8.13580194e-01]
 [ 2.73088381e+00  4.94208622e-01  2.40912410e+01 -9.78833361e-01]
 [ 4.42976173e+00  1.32587598e+00  2.14078390e+01 -1.36236943e-01]
 [ 1.99883773e-01  2.23000249e+01  1.02007951e+00  6.63888874e-01]
 [ 4.55693934e+00  9.37724341e-01  2.21449704e+01  2.13559375e+00]
 [-4.96532651e-01  2.26651211e+01  7.66825731e-01  1.50431622e+00]
 [ 2.12281289e+01  4.18225865e+00 -5.23206228e-02  1.34405931e+00]
 [ 2.14646252e+01  1.09255203e+00 -3.62409531e-01  4.94317264e+00]
 [ 1.26896647e+00  5.13390523e+00  2.18674015e+01  1.50510621e-02]
 [-6.46223517e-01  2.11783009e+01 -1.15102786e+00  7.96217052e-01]
 [ 2.43313678e+01  9.71620044e-01  3.65769570e-02  9.96512390e-01]]
```

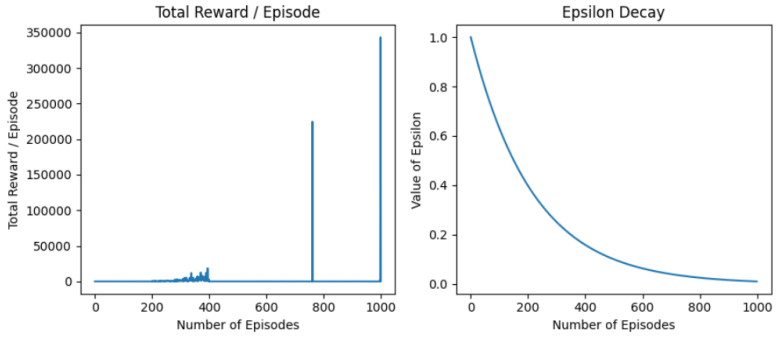
```
Trained Q-table_A:
[[ 2.37403482e+01 -3.43253854e-01  4.46029039e-01  1.92070475e+00]
 [ 4.65384867e-01  7.36041192e+00  2.23697815e+01  2.85232175e-01]
 [ 1.81742353e+00 -8.61542185e-02  2.21963820e+01  2.02727599e-02]
 [ 2.22027823e+01  1.40392428e-01  6.75540232e-01 -3.83484844e-01]
 [ 1.41233053e+00  2.26937790e+01  3.99534457e-02  7.82866433e-01]
 [ 2.17972880e+01 -4.74753740e-01 -4.10861537e-01  7.08906680e-01]
 [-5.66337908e-01  7.42687670e-01  2.12329619e+01 -7.82993990e-02]
 [ 1.55507755e+00  1.76618927e+00  2.19127581e+01  7.52527356e-01]
 [-2.37509346e+00  2.37952168e+01 -5.59406351e-01  7.93430423e-01]
 [ 3.75624537e+00 -6.34597555e-01  2.50531852e+01  3.18521063e+00]
 [ 4.69221965e-01  2.25166494e+01  2.07114387e+00  6.38709775e-01]
 [ 2.26864463e+01  2.28787183e-01 -1.28691343e+00 -9.20537456e-01]
 [ 2.54666865e+01  2.85247994e-01  5.60030441e+00  2.36807459e+00]
 [ 1.70513950e+00 -3.81277170e-02  2.05136084e+01  8.92841489e-01]
 [ 3.85461958e-01  2.29426740e+01  2.69729793e+00 -7.23426594e-01]
 [ 2.33671658e+01 -2.20474945e-01  1.85525450e+00  2.84282000e+00]]

Trained Q-table_B:
[[ 2.46100794e+01  6.01217661e+00  7.38397274e-01  4.01725355e+00]
 [ 2.37231739e+00  3.15278850e-01  2.29476274e+01 -5.13881295e-01]
 [ 1.26374428e+00  8.23089434e-01  2.34073089e+01 -4.17153945e-01]
 [ 2.19418104e+01  1.82340132e+00  8.46033828e+00 -3.24926756e-01]
 [-2.97684612e+00  2.20912556e+01  5.28325404e+00  8.26539403e-01]
 [ 2.47734581e+01  2.07905339e+00  4.48766687e+00  8.13580194e-01]
 [ 2.73088381e+00  4.94208622e-01  2.40912410e+01 -9.78833361e-01]
 [ 4.42976173e+00  1.32587598e+00  2.14078390e+01 -1.36236943e-01]
 [ 1.99883773e-01  2.23000249e+01  1.02007951e+00  6.63888874e-01]
 [ 4.55693934e+00  9.37724341e-01  2.21449704e+01  2.13559375e+00]
 [-4.96532651e-01  2.26651211e+01  7.66825731e-01  1.50431622e+00]
 [ 2.12281289e+01  4.18225865e+00 -5.23206228e-02  1.34405931e+00]
 [ 2.14646252e+01  1.09255203e+00 -3.62409531e-01  4.94317264e+00]
 [ 1.26896647e+00  5.13390523e+00  2.18674015e+01  1.50510621e-02]
 [-6.46223517e-01  2.11783009e+01 -1.15102786e+00  7.96217052e-01]
 [ 2.43313678e+01  9.71620044e-01  3.65769570e-02  9.96512390e-01]]
```

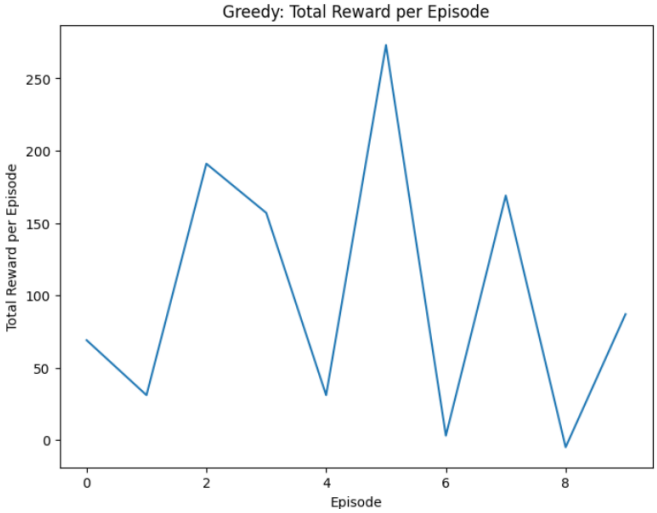
Hyperparameter Tuning - Case 6:

Number of episodes: 1000

Epsilon decay: 0.9954

In this case I increased the number of episodes again, this time to 1000. It didn't change much in the non-greedy one, but the change is reflected in the greedy one as it learnt more.





```
Q-table_A: At the beginning
[[193.89598465 188.96382132 191.39926035 195.52431082]
 [195.50106446 193.59166926 190.95432416 195.43182628]
 [195.26666098 191.35660023 195.47394796 193.49638517]
 [192.25098617 194.13867741 191.5526877  194.4175862 ]
 [193.4997827  190.65158812 196.11977151 195.19395833]
 [192.72879083 192.16339658 191.21883106 194.83364753]
 [192.22608375 189.2487419  190.71652824 193.12461767]
 [195.20357527 183.11015589 191.79365279 189.60567352]
 [193.27806199 181.89440012 190.47840154 186.29355394]
 [194.1841539  192.39095602 191.1256664  194.57120601]
 [198.3960178  195.31536908 194.38977271 194.49487746]
 [187.84790509 192.32317214 194.20383398 194.24714021]
 [179.89110708 192.84526926 187.14641731 194.00496923]
 [190.16531543 194.68164041 192.45369657 194.83510984]
 [194.96964843 194.70112982 195.69836919 194.90527555]
 [192.64508396 194.5567609  193.23978427 194.78560614]]

Q-table_B: At the beginning
[[196.23563009 188.17334717 192.94480761 194.61919173]
 [193.44552732 191.84835679 186.99285652 193.62152457]
 [192.56258034 195.77058048 187.91811362 194.02670925]
 [188.69501335 185.12036383 194.51331004 193.6180525 ]
 [194.88835499 186.02780512 192.78936206 193.59613558]
 [194.42224712 194.02265181 191.27775155 190.52912289]
 [195.25074856 193.26163711 196.17143807 194.07656511]
 [195.01472648 192.56499444 194.03771327 197.54590135]
 [190.81691146 191.67902078 194.09043061 196.38064183]
 [193.82203239 195.59370634 195.59888568 193.34292728]
 [190.42584984 192.71156561 192.9695761  193.80435152]
 [192.09635333 192.35473329 191.96260161 191.05518266]
 [190.96135123 190.33891537 184.14984427 193.64346243]
 [191.14534838 190.19124416 192.40316987 195.50163522]
 [193.48782999 192.59659319 190.77792332 193.58921203]
 [193.73306483 191.72296297 190.85660795 192.68704503]]
```

```
Trained Q-table_A:
[[193.89598465 188.96382132 191.39926035 195.52431082]
 [195.50106446 193.59166926 190.95432416 195.43182628]
 [195.26666098 191.35660023 195.47394796 193.49638517]
 [192.25098617 194.13867741 191.5526877  194.4175862 ]
 [193.4997827  190.65158812 196.11977151 195.19395833]
 [192.72879083 192.16339658 191.21883106 194.83364753]
 [192.22608375 189.2487419  190.71652824 193.12461767]
 [195.20357527 183.11015589 191.79365279 189.60567352]
 [193.27806199 181.89440012 190.47840154 186.29355394]
 [194.1841539  192.39095602 191.1256664  194.57120601]
 [198.3960178  195.31536908 194.38977271 194.49487746]
 [187.84790509 192.32317214 194.20383398 194.24714021]
 [179.89110708 192.84526926 187.14641731 194.00496923]
 [190.16531543 194.68164041 192.45369657 194.83510984]
 [194.96964843 194.70112982 195.69836919 194.90527555]
 [192.64508396 194.5567609  193.23978427 194.78560614]]

Trained Q-table_B:
[[196.23563009 188.17334717 192.94480761 194.61919173]
 [193.44552732 191.84835679 186.99285652 193.62152457]
 [192.56258034 195.77058048 187.91811362 194.02670925]
 [188.69501335 185.12036383 194.51331004 193.6180525 ]
 [194.88835499 186.02780512 192.78936206 193.59613558]
 [194.42224712 194.02265181 191.27775155 190.52912289]
 [195.25074856 193.26163711 196.17143807 194.07656511]
 [195.01472648 192.56499444 194.03771327 197.54590135]
 [190.81691146 191.67902078 194.09043061 196.38064183]
 [193.82203239 195.59370634 195.59888568 193.34292728]
 [190.42584984 192.71156561 192.9695761  193.80435152]
 [192.09635333 192.35473329 191.96260161 191.05518266]
 [190.96135123 190.33891537 184.14984427 193.64346243]
 [191.14534838 190.19124416 192.40316987 195.50163522]
 [193.48782999 192.59659319 190.77792332 193.58921203]
 [193.73306483 191.72296297 190.85660795 192.68704503]]
```

Ans to ques 4:

Compared the performance of SARSA and Double Q-Learning on the same environment with eplison decay of 0.0995 and 100 attempts for both.

I can see that in the same environment both of the algorithms are performing pretty well and similarly. Although SARSA seems to be a little bit stable and converging slowly. However, the difference is not much noticeable in terms of performance.