Week 5 lab is a demonstration of TD(0) and Q-Learning method. Similar to Monte Carlo method, TD(0) predict state value without prior knowledge. However, MC calculate all states values in the policy in the iteration while TD(0) gives the value of the state immediately after calculating it before arriving to the terminal state (distention), which is faster, and more efficient in some applications such as traffic prediction. In the file, states values are discovered after 1000 iteration, which I think it is too many iterations for 4\*3 grid with explore/exploit ratio of 0.1, I changed that to 100 and still TD(0) gave all states values. I also tried 50 and there is one states needed to be discovered. There has to be a way to calculate the needed number of iterations based on learning ratio, given policy and size of the grid, and action related to each state. The function picks a random action and move to the new state and calculate which happens 10% of the total iteration because the learning ratio is 0.1, the 90% if following the policy, the policy will not go over all states which make the function explore a new state 10% of the iteration. So, I am not sure how we can figure out the minimum number of iterations needed to find all states values.

Q-Learning is similar to TD(0). Q-Learning seeks to find best action for a state. It does not require a policy because Q-learning seeks to find the optimal policy that gives maximum reward. The way it works is building a Q table where it contains all possible reward for all actions of all state. It builds the table by starting at start state and pick an action with probability epsilon and get the reward for the state for specific action and put it in the table and it continues to build it the table in that way. Then, it extracts the optimal policy by taking the best action of the state. I was wondering if the exploration rate will change anything to the extracted policy, I changed epsilon from 0.5 to 0.1, it did not change anything. However, it took more time than 0.05. Also, changing the Alpha changed the final policy entirely.