Assignment 1

Task 1

Max Reward: 56.0 Average Reward: 21.01 Standard Deviation: 9.43

Task 2

Max Reward: 85.0 Average Reward: 36.8 Standard Deviation: 15.16

The max and average reward from the deterministic policy in the cartpole enviroment are higher than the random policy, showing the value of a deterministic policy over a random for this exercise.

Task 3

The RecordVideo and RecordEpisodeStatistics were used to capture each episode as well as the statistics.

Task 4

4a: Max Reward: -5.62 Average Reward: -6.61 Standard Deviation: 0.42

4b: Max Reward: 96.22 Average Reward: 82.83 Standard Deviation: 33.01

In the mountain car environment, a car is placed in the middle of a valley. The objective is to push the car to the top of the right side and reach the goal. The observation space of this environment contains the position of the car, and the velocity of the car. If the velocity of the car is negative then the car is traveling left. If the velocity is positive, then the car is traveling right. Reaching the goal is determined by the position of the car being greater or equal to 0.45.

For both the random and deterministic policies, the episode were limited to 200 steps.

The results of the random policy show that the car never reaches the top of the mountain in those 200 steps. The max and average reward are negative. This is because a positive reward is only received when reaching the goal.

The deterministic policy chosen is to push the car in the direction of its current velocity. The idea is to build enough momentum by oscillating left and right until enough momentum is built to reach the top. This policy does much better

at maximizing the reward, with the car consistently reaching the goal, and receiving a positive reward.

Task 5

An example of a real world decision task is stock trading. You have a portfolio of stocks and their prices, trading volumes, number of shares, and other pieces of data. This data can be viewed as the environment in the case of re-enforcement learning. You have a discrete action space for each stock: buy shares, sell shares, hold. And there is a clear reward based on the change in portfolio value. In the persepective of this assignment, a random policy would not work well. You could design a rudimentary deterministic policy where you buy shares of a stock when the price is below its historical average, and sell shares when it is above its historical average.