CS 440 MP1 Part2
Dhruv Vajpeyi

1.

(a) The total reward received by the robot starts negative and then is always 0.0. It does not attempt to deliver any packages.
In WorldWithoutThief, the bot has to pass through slippery patches to deliver the packages. When it slips on the patch, it receives a negative reward(penalty). Since there is no randomness, it will stop passing through the slippery patch and will therefore deliver no packages.

(b) The total reward received per episode is low(-20 to low tens) for the first 25 or so episodes, then climbs to upper hundreds/lower two hundreds, and then stays at that level with some variation for the rest of the episodes.
The bot does better in this simulation because it still has a random chance to pass through the slippery patch and deliver the package, thus gaining a net reward. Because of the reward it receives, its policy is updated to pass through the patch.

(c) The total reward received per episode is very low.(Mostly negative).
The bot does terribly in this simulation because it makes too many random choices and therefore does not learn any optimum path. An ε value of 1 would be completely random.

2.

(a) The total reward received by the robot per episode is always poor(Usually negative).
The agent cannot formulate an optimal policy because the state does not take into consideration the position of the thief. Therefore, all actions are made independent of the thief and the agent gets caught a lot.

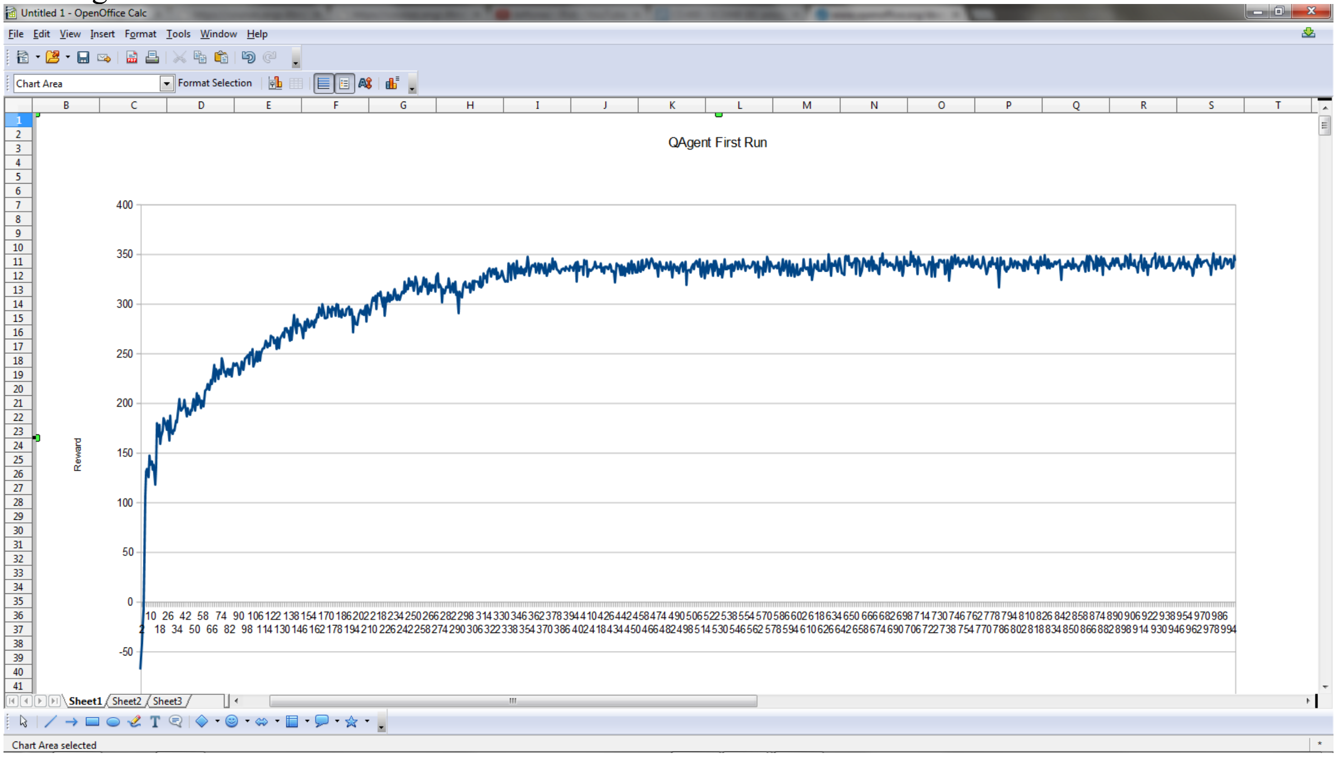
(b) The total reward received per episode grows quickly and remains very high.(Upper 200s).
Since the state takes the thief's position into consideration, the agent's policy causes it to avoid the thief. Also, the agent does not need to risk traveling through slippery patches in WorldWithThief.

(c) Starting with epsilon = 0.1, learning rate = 0, Average reward =-103
learning rate = 0.1, Average reward = 210
epsilon = 0.2, Average reward = 95
epsilon = 0.05, Average reward = 280
epsilon = 0.025, Average reward = 311
epsilon = 0.0125, Average reward = 331
epsilon = 0.006, Average reward = 336
epsilon = 0.003, Average reward = 316
epsilon = 0.005, Average reward = 340
learning rate = 0.075, Average reward = 320
learning rate = 0.125, Average reward = 330.

Optimal: epsilon = 0.005, learning rate = 0.1

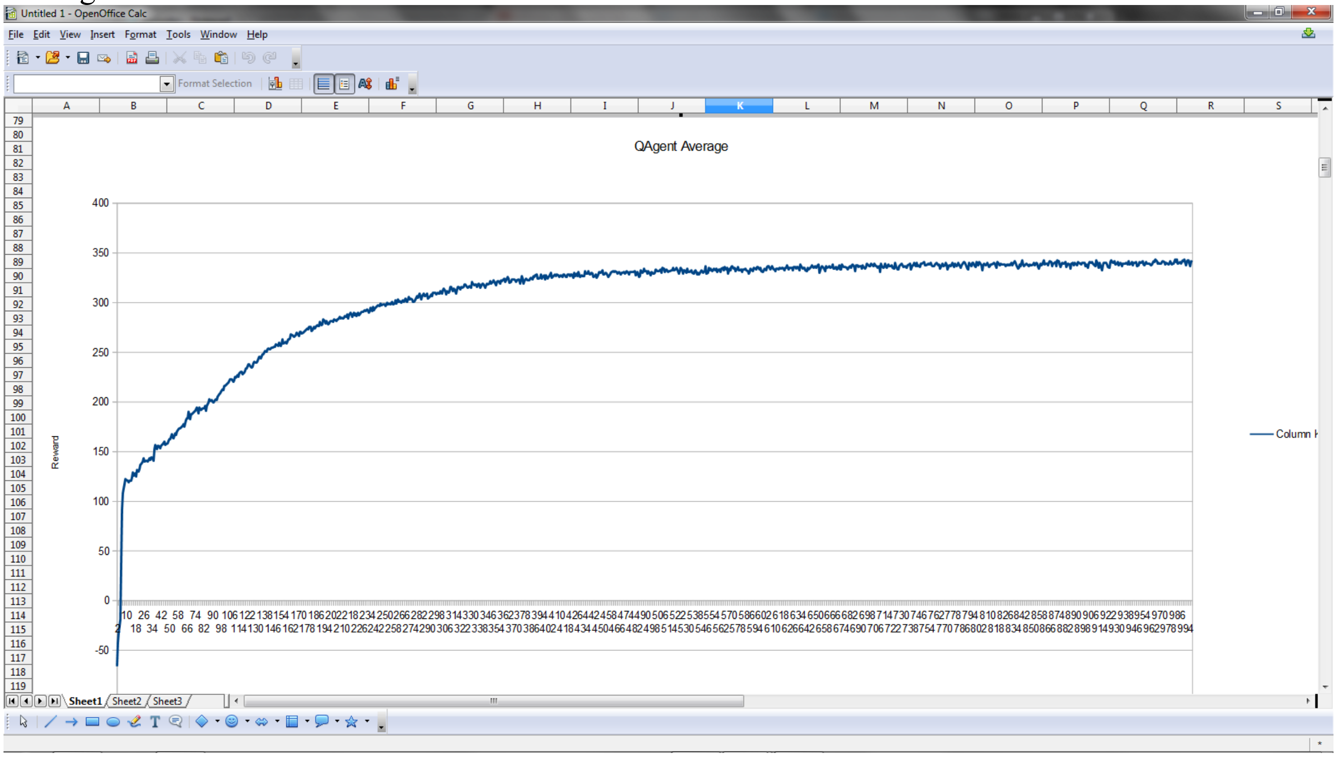
3.

First Agent



The reward starts negative but quickly increases to around 100, and then gradually increases upto 350. Throughout the graph there is slight variation between episodes.

Average of 10



The graph is similar but there is less variation between adjacent episodes.