**Alexander Khrulev 500882732**
**Mahan Pandey 500881861**

Part 1: UCB algorithm

Our results provided in the graphs below show that the exploration constant C impacts average rewards over lifetime by changing how much the agent explores vs exploits.

Comparing C=0.01 and C=0.1 we see that C=0.01 has the most runs with the highest average rewards but also has many outliers compared to C=0.1. Looking at the graph for C=0.01 we can see the 3 lowest runs grey, blue and brown where the average reward is around or below 60%. Since the exploration constant C is low, the agent chooses the action that maximizes Q without considering if Q is an accurate estimate of the true values q. For these runs we can say the agent is exploiting too much and not exploring enough.

As C increases, we see runs with a wide range of average rewards over lifetime.

For C=1 we see it has the lowest number runs with high average rewards over lifetime because it explores too much and does not exploit what it has found to be the optimal action. Since C is high, even after finding the true optimal action the agent will try new actions that haven't been tried as much.

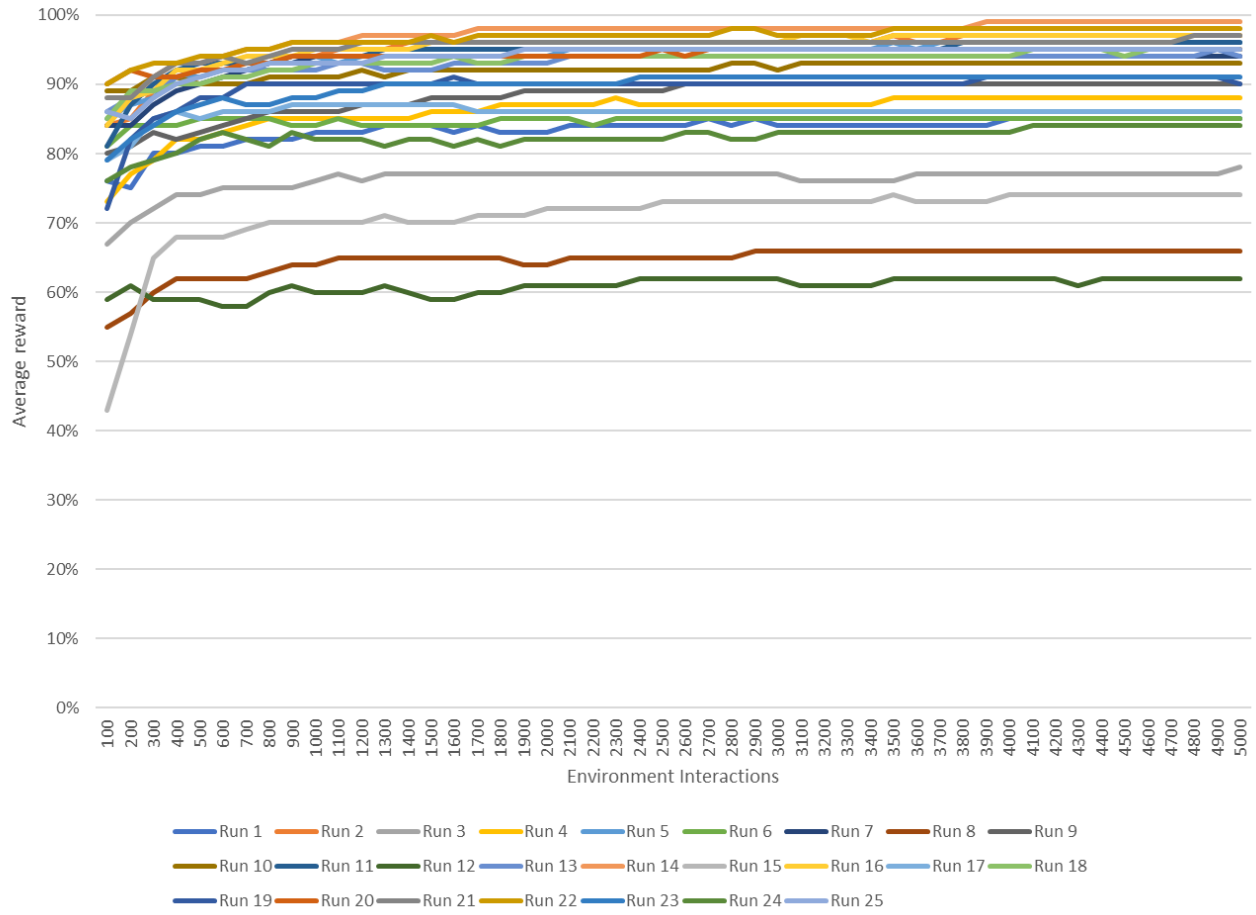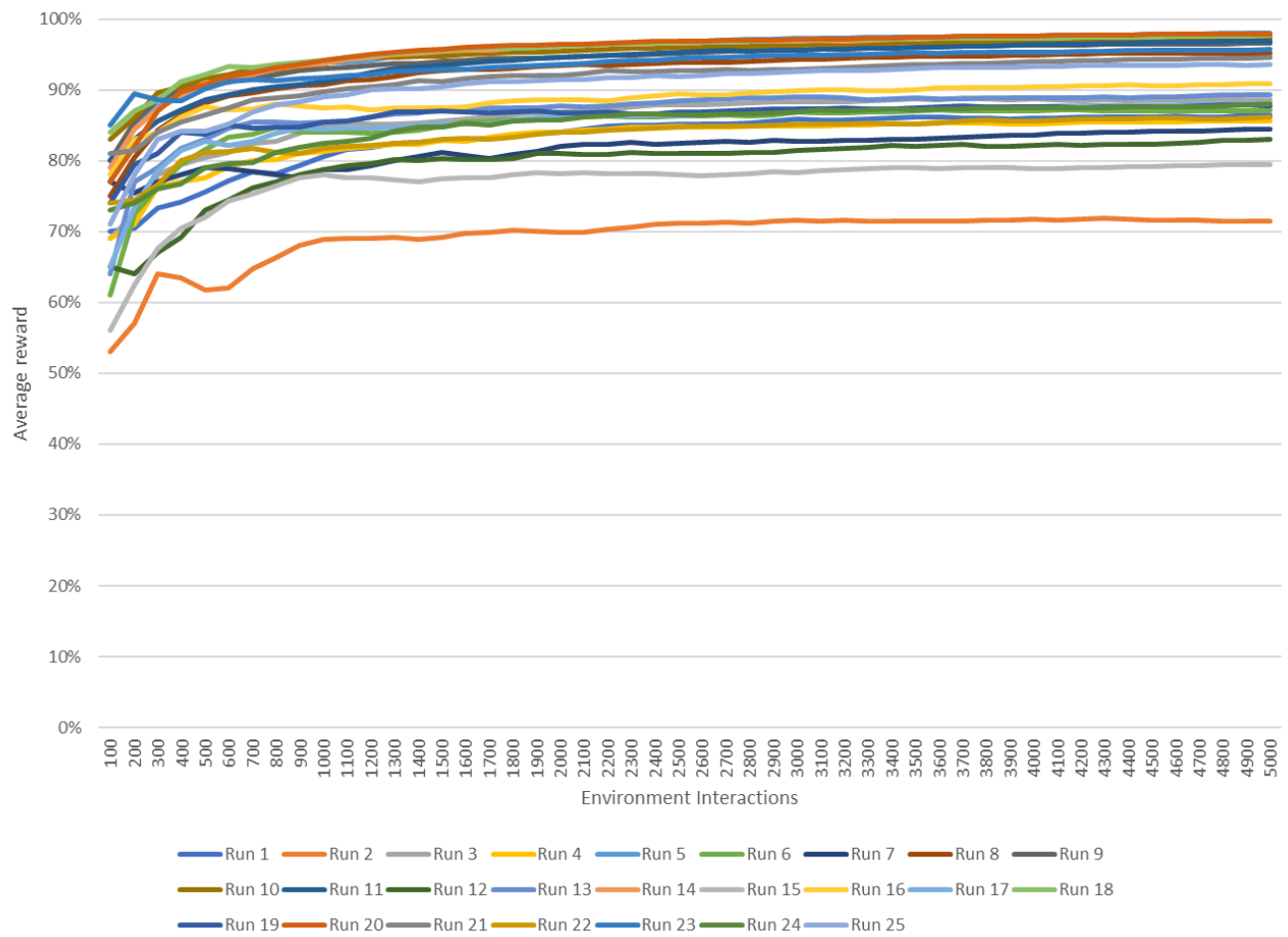Environment Interactions vs Average Reward C=0.1
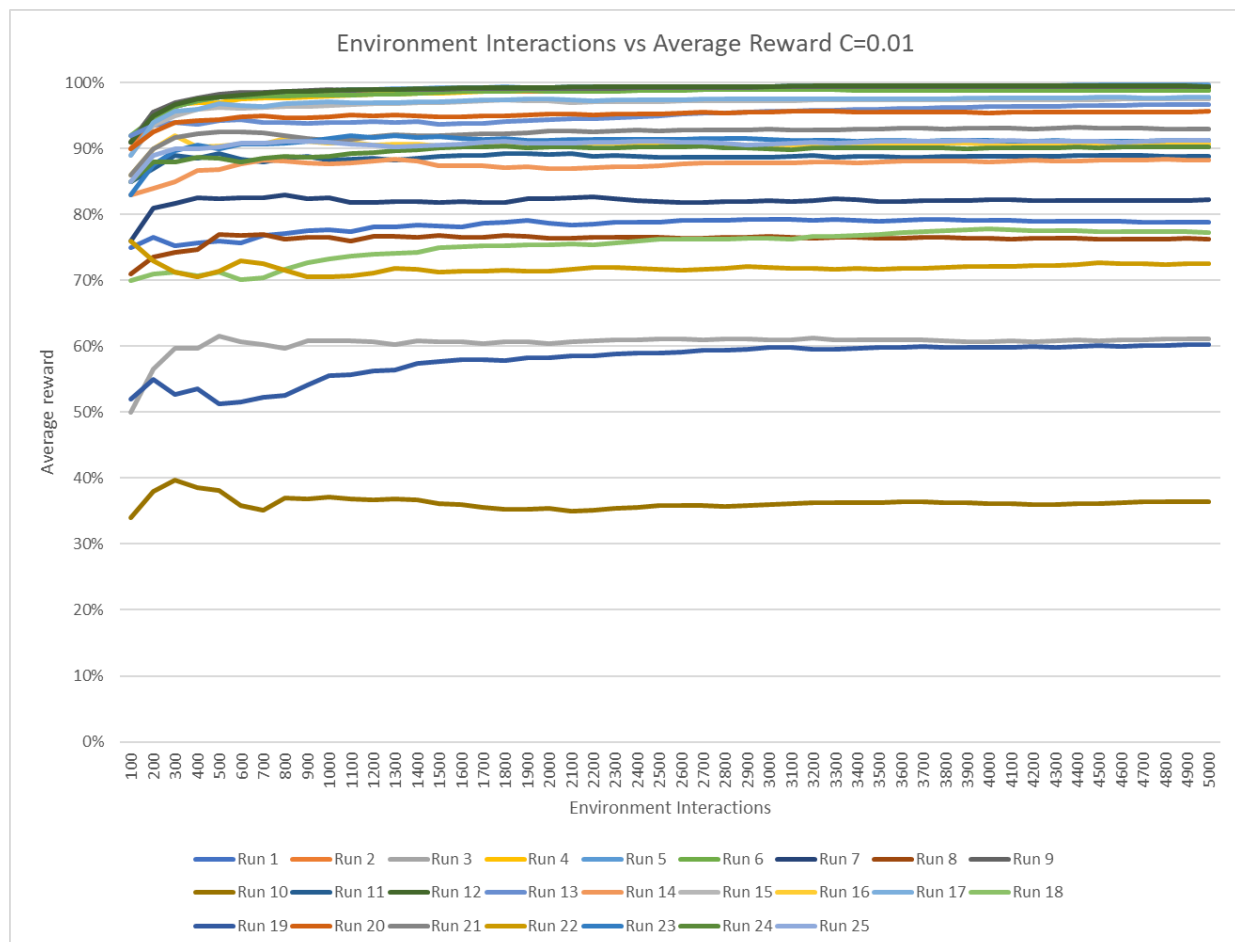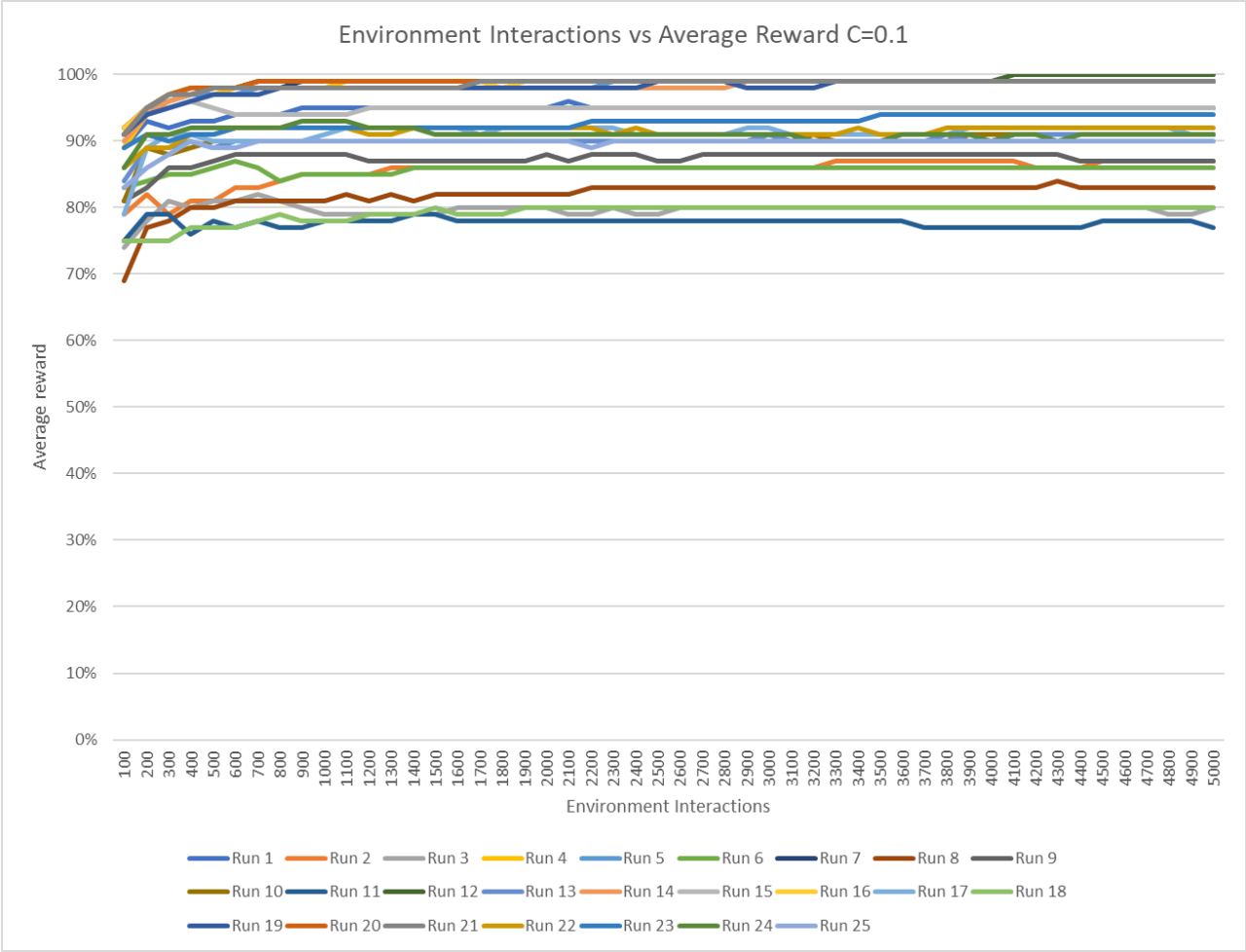
Environment Interactions vs Average Reward C=0.25
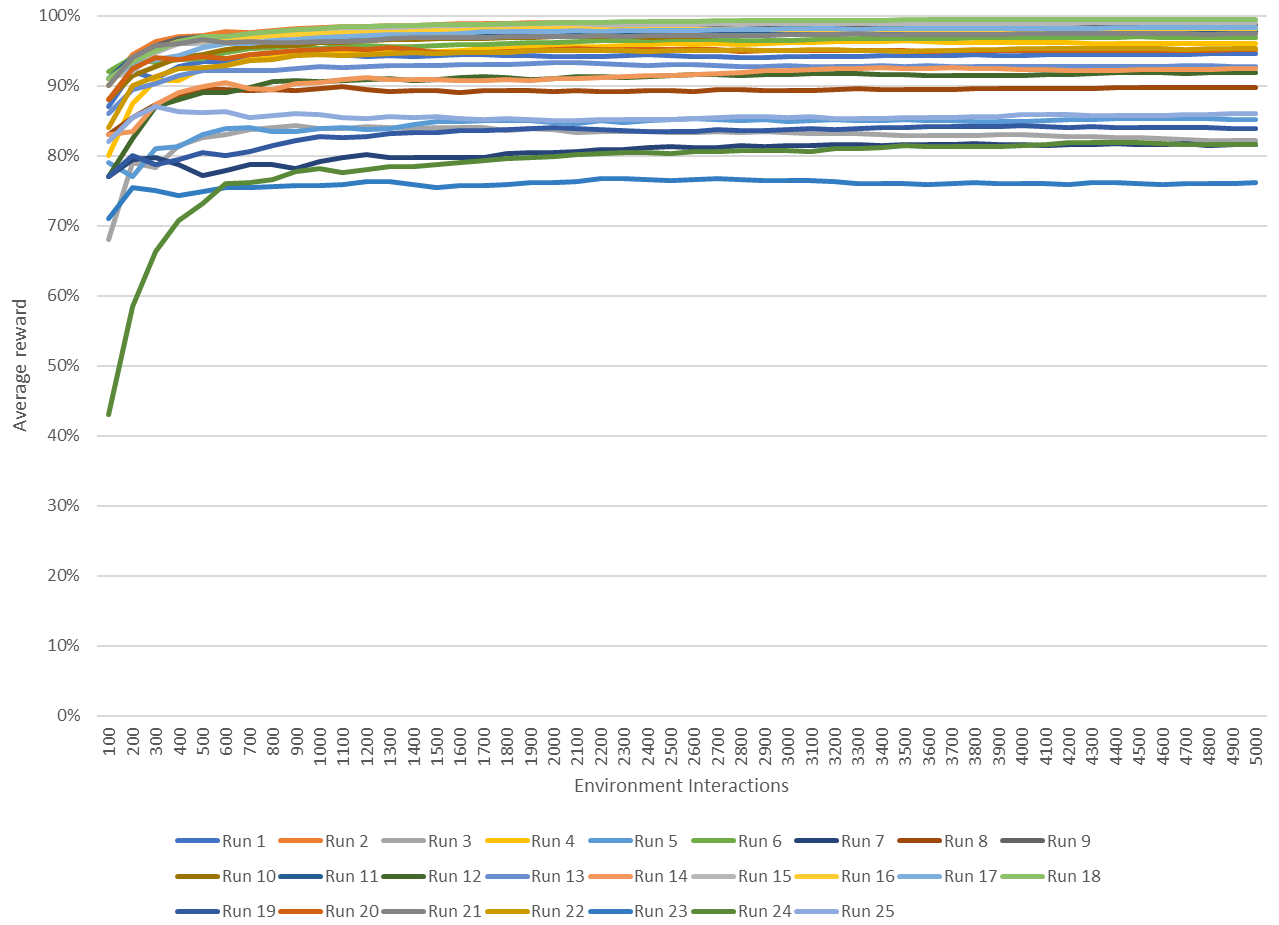
Environment Interactions vs Average Reward C=0.5

Environment Interactions vs Average Reward C=0.75

Environment Interactions vs Average Reward C=1