For week 1 lab, we are learning about multi-armed bandit model. Lab 1 contains 2 important files; the first file explains how single armed bandit works. Basically, single armed bandit gives a sequential random number, it gives a random number each time we pull the “arm”, to calculate the mean of the output of the single armed bandit each iteration, which makes the mean stabilizes on the true mean over long time, which I noticed. Also, in my opinion, I think when the mean stabilizes it means that exploration phase is no longer needed. However, we do not have another bandit to explore or exploit currently. When I run the experiment with other attributes for the size of sample and the sample mean and variation, I noticed that the mean, and standard deviation are heading toward the true value that they’ve been giving after 10% to 20% of overall iterations.

As for the second file which interduce the epsilon greedy policy, this policy is to select a bandit in each iteration based on the maximum mean of a bandit, to gain a knowledge of the mean of a bandit, it needs to explore all bandit and build a knowledge so it can exploit a bandit. Also, after a certain time it needs to re-explore, and this is because the mean change each iteration and it could exploit the wrong bandit that gave a high mean at the beginning, but other bandits are higher after a while. The epsilon greedy policy will not give the maximum number of rewards because it has to build a knowledge of which bandit that has the highest reward. Epsilon greedy policy also can be improved to get more total rewards by changing the exploration phase occurs. We do not want more frequent exploration because we need to exploit the highest bandit at that time, and we do not want less frequent exploration because we need to update the means of each bandit