Q1:

1. In the initial episodes, both Sarsa and Q-learning outperform MC-Control, learning more rapidly. However, MC-Control catches up as the number of episodes increases, reflecting its effective exploration and learning from complete sequences. Post-100 episodes, the learning curves converge, indicating a plateau in performance improvement. Towards the final episodes, Q-learning achieves a higher average return, indicating a slightly more effective policy.
3. Yes they do as:
   1. Q-learning is an off-policy and Sarsa is an On-Policy
   2. This means that even though both learn their plocicy by exploration
   3. Q-learning actualy learns two policies, a behavioural and an optimal policy
   4. The bahavioural policy is only used to generate experience
   5. While the target policy is the oen we want to actual learn#
   6. For sarsa, the policy learned is near optimal
   7. Meanwhile for Q-learning it is the optimal policy