Question 1)	
-	9n the step, & (St, At) ← Average (Returns (St, At))
	we are taking arresage every time and thus
-UT 161 91	we are taking average every time and thus it vivolves a lot of calculations
	we can avoid this by just stooing the value of
	of and the number of times the particular state action pair has been encountered.
	state artière noix has been encountered.
	state adon pass mos
**	1.e.
	R(GLAH) < 01GLAL) + 1 LG1-B(SBAL)7
	$g(s_{\bullet}, A_{\bullet}) \leftarrow g(s_{\bullet}, A_{\bullet}) + 1 \qquad g(s_{\bullet}, A_{\bullet})^2$ $count[s][a]$
	lseudo code:
	Initialize:
	TT(S) & A(S)
	Q(5,a) E IR
0-	Count 16 07 - 8 + 5, a
	Loop forever (for each episode) (leutour one pure) !
	GUITTE Chanse SOES, AO EA(SO)
do.	monerate an episode from so, to tollowing !!
20 40	So, Ao, R1,, ST-1, AT-1, RT
And The Co	Inizionalia (u.e) sub art
*	Loop for each step of episade, t=T-1, T-2,, 0:
	Gre YG1 + KE+1
and the same	Count [St][At] +=)
	Unless the paid St, At appears in So, Ao, Si,, Sti, Atri
`	and of the box ketter
	$g(s_{t}, A_{t}) \leftarrow g(s_{t}, A_{t}) + \underbrace{I}_{\text{count}(s_{t}, A_{t})} $
F.	
**:	$\pi(s_t) \leftarrow \operatorname{argmax} g(s_t, a)$

Question 2: In Monte Casto Es, y we use q(2,a) pair at the start of opisode and continue till we reach a terminal state. This can be shown in the form of bockup diagram as follows. Later Court of Terminal state. We need to martain an assay of time steps Guestion 3 for the pair (5, a), which will store the time instents when that pair is visited for that [r episode. Updated formula internes of Q13, a) can be written as: g(s, a) = Etet(s,a) Privated & P: 4th-1

Lev(s,a) Lan District tersa)

6	
(=	where
	1(t) - first time of termination following time to
(-	$T(t)' \rightarrow fisst$ time of termination following time to $G_{1t} \rightarrow s$ return after t up through $T(t)$
	grow of woody (CC)
Que	
Sue	shions
	We move to a new building and in which
	we move to a new building and
	Lot, which is near the same highway and we have previous experience of driving to the old building
	previous experience al disvoca / Manay and we have
	This is because In this so the old building.
	previous experience of driving to the old building. This is because: In this case, only a part of the
	CINC DIVIDITATION OF ALL
_	Now, in the 7D method the state values are updated
	full polenda the without the need to generate the
	on the fly i.e. without the need to generate the full episode. Thus since the starting values for these states are updated
	States are already close to true values it will lead to faster convergence.
	that to faster winvergence,
	Tes same thing will happen in on ginal scenario because of the above reason if initial state values estimate is close to tope values
	of the above season if initial state alues estimate is close
=	to true values,
-	port of the state
Question8.	In case of greedy action selection Q-reasoning
	In case of greedy action selection B-learning will not be the same as SARSA.
12	This is because the way we note a in tall
	This is because the way we update & is both the methods differs. In Q-learning the q values
	and world increase that all is a line of values
	are updated irrespective of the action taken and
r .	the action is the chosen as per the updated & value.
	Whereas in SARSA next action is chosen according to
S.	the action is then chosen as pex the updated & value. Whereas in SARSA next action is chosen according to the current & value and then & is updated.
1	ov, silve the house the house of the land of the terms of the
7	is both cases, both the methods perform differently.
	merrious peoposition affectently.