QUESTION: *Observe what you see with the agent's behavior as it takes random actions. Does the**smartcab** eventually make it to the destination? Are there any other interesting observations to note?*

By choosing random actions, agent regardless to environments rulles and the outcome of his previous actions, is taking some actions and sometimes it reaches the destination but it has no strategy to find the destination.

QUESTION: *What states have you identified that are appropriate for modeling the **smartcab** and environment? Why do you believe each of these states to be appropriate for this problem?*

Based on traffic rulles we must have appropriate understanding from oncoming and left vehicles and also traffic light status but right vehicle is not necessery since having other features and keeping in mind that all vehicles are following the traffic rulles then we can dismiss right status. We also need to add next_waypoint to our status because agent should decide the route based on GPS data and consider other included stats to learn the best policy. For agent to learn how to obey traffic rulles and find the path way, having 'deadline' feature is not necessary and it is just a factor to evaluate agent's performance. If we add this feature it will expand states space by the number of deadline let's say now we have 128 states and deadline is 100 then the state will expand to 12800 !

OPTIONAL: *How many states in total exist for the **smartcab** in this environment? Does this number seem reasonable given that the goal of Q-Learning is to learn and make informed decisions about each state? Why or why not?*

There are 4 options for 3 features ("oncoming", "left", "next_waypoint") and 1 option for "light" therfore maximum 128 states should exist. This number is not significantly high and as our experiment later shows agent can learn a relatively good policy in the very first trials.

QUESTION: *What changes do you notice in the agent's behavior when compared to the basic driving agent when random actions were always taken? Why is this behavior occurring?*

In the first few trials agents still has some random actions which do not lead to the desination but after some steps agent starts to follow the right path and finally reaches the destination. This bedavior is reasonable becuase at first agent has know understanding about environment and the states but after some trial and errors and getting the negative and positive rewards, agent learns the right actions in each state.

__QUESTION:__ Report the different values for the parameters tuned in your basic implementation of Q-Learning. For which set of parameters does the agent perform best? How well does the final driving agent perform?

Epsilon	Gamma	Alpha	Success Rate (Winning Numbers)
0.05	0.7	0.2	78
0.05	0.7	0.3	82
0.05	0.8	0.2	80
0.05	0.8	0.3	56
0.06	0.7	0.2	71
0.06	0.7	0.3	65
0.06	0.8	0.2	64
0.06	0.8	0.3	48

Table 1.1 (Q-Learning success rate report based on different parameters)

As being shown in Table 1.1 we can see that the parameters set (0.05, 0.7, 0.3) is giving the best learning result.

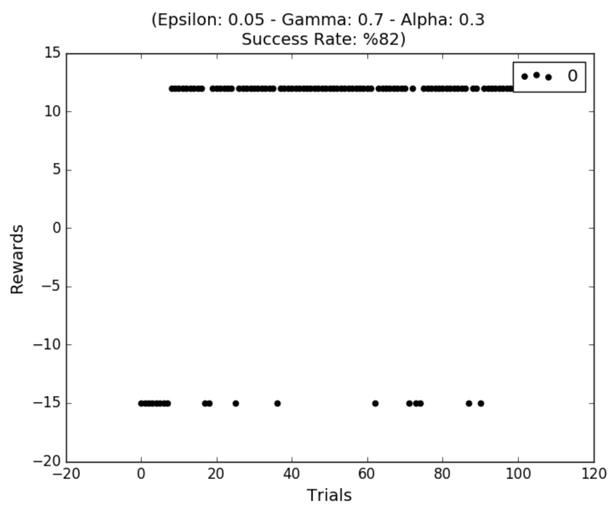

Figure 1.1 (Success rate scatter plot)

We can see from Figure 1.1 that almost after 20th trial, agent is able to reach the destination with rewards of +12. Agent has a success rate of 82%. (to cmpare this result to other sets of parameters you can refer to appendix part one)

QUESTION: *Does your agent get close to finding an optimal policy, i.e. reach the destination in the minimum possible time, and not incur any penalties? How would you describe an optimal policy for this problem?*

Optimal policy is the one that helps agent to reach the destination in minimum time and violation with keeping in mind that this is really related to the agents distance to destination and traffic condition. These two factors may affect agent to behaves differently in diffrent situation. Therefore I have measured the policy by the proportion of positive rewards toward negative ones. Figure 1.2 shows the agents positive rewards proportion for the best set of parameters discussed above.

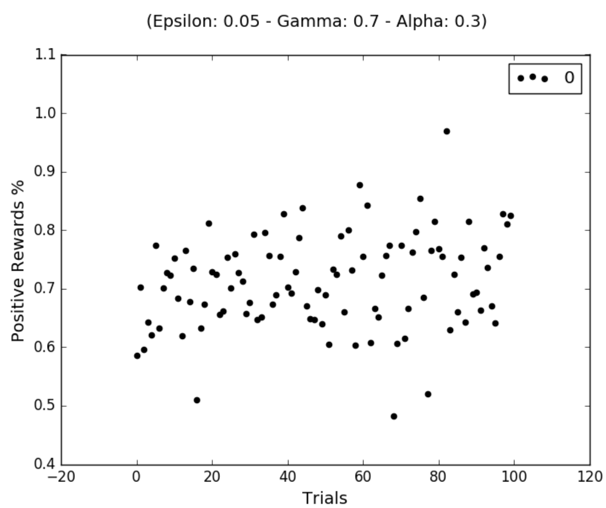

Figure 1.2 ((positive rewards counts) /(positive + negative rewards counts))

In Figure 1.2 we can see that by time passing, agent is reaching a better policy by getting more positive rewards than negative ones. This policy is not 100% optimal but in last 10 trial it reaches the destination with 60% to 90% getting positive rewards.

Appendix – 1 (Success Rates)