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?

The smartcab doesn't reach its destination most of the time (almost all the time in my testing). It takes random action choices and the cab moves slowly away from the start point in a random direction.

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?

I found combination of inputs be appropriate for capturing state information as shown below:
self.state = (self.next_waypoint, inputs['light'], inputs['oncoming'], inputs['left'], inputs['right'])
Using above state combination, we can retrieve best action based on rewards captured for (state, action) pairs.
self.next_waypoint – captures the best route planner action
inputs['light'] – capture the traffic light at the intersection
inputs['right','left','incoming'] – capture the oncoming traffic at the intersection

Rewards are based on the best possible taken by smartcab at the intersection. We need this state information so it can learn the rewards it gets from the environment based on previous actions taken and continue to take best possible action through Q-Learning.

The other possible variable that can be included in state is self.env.get_deadline(self), by including this variable, learning takes lot longer since the number of possible states increases, and Q-learning takes longer.

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?

Total possible states for below state:
```
self.state = (self.next_waypoint, inputs['light'], inputs['oncoming'], inputs['left'], inputs['right'])
```
Total possible = 3*2*4*4*4

This is appropriate for current learning because of the short grid size and small number of deadline iterations that we take. It takes longer to train with the increase in states.

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?

The smartcab moves towards destination because of Q-Learning, it chooses the best action based on the current state of the environment and also learns the destination route through accumulative rewards. The smartcab tries to choose next step by evaluating Q value for all of its immediate (state,action) pairs and updates the rewards that it receives after taking that action, by repeating these steps, it will learn the best possible waypoint by following traffic rules to reach the destination.

The smartcab reached destination for most of the time with Q-Learning. The smartcab initially chooses its action randomly as it receives feedback from the environment and as it learns the environment, it seems to take right steps towards the destination. The initial steps seem random and doesn't seem to get stuck in local minima once it learns the environment.

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?

Following are the parameters tuned:
self.gamma(discount factor) and self.alpa(learning factor)

For 1000 trails:

gamma=0.10, alpa =0.10, epsilon=0.10: total_steps= 13238, net_rewards = 22369.0, negative_rewards = -69.0, successful_trials = 998
gamma=0.10, alpa =0.10, epsilon=1.00: total_steps= 12681, net_rewards = 22107.5, negative_rewards = -66.5, successful_trials = 998
gamma=0.10, alpa =0.33, epsilon=0.10: total_steps= 13216, net_rewards = 22357.0, negative_rewards = -82.5, successful_trials = 999
gamma=0.10, alpa =0.33, epsilon=1.00: total_steps= 12962, net_rewards = 22330.0, negative_rewards = -72.0, successful_trials = 997
gamma=0.10, alpa =0.55, epsilon=0.10: total_steps= 12580, net_rewards = 22355.0, negative_rewards = -69.0, successful_trials = 997
gamma=0.10, alpa =0.55, epsilon=1.00: total_steps= 13707, net_rewards = 22316.0, negative_rewards = -74.0, successful_trials = 997
gamma=0.10, alpa =0.78, epsilon=0.10: total_steps= 12802, net_rewards = 22346.5, negative_rewards = -66.5, successful_trials = 998
gamma=0.10, alpa =0.78, epsilon=1.00: total_steps= 12897, net_rewards = 22172.0, negative_rewards = -72.0, successful_trials = 996
gamma=0.10, alpa =1.00, epsilon=0.10: total_steps= 12539, net_rewards = 22405.5, negative_rewards = -74.0, successful_trials = 997
gamma=0.10, alpa =1.00, epsilon=1.00: total_steps= 12793, net_rewards = 22250.0, negative_rewards = -74.0, successful_trials = 997
gamma=0.33, alpa =0.10, epsilon=0.10: total_steps= 13344, net_rewards = 22387.5, negative_rewards = -134.0, successful_trials = 994
gamma=0.33, alpa =0.10, epsilon=1.00: total_steps= 13564, net_rewards = 22801.5, negative_rewards = -299.0, successful_trials = 981
gamma=0.33, alpa =0.33, epsilon=0.10: total_steps= 13230, net_rewards = 22332.0, negative_rewards = -98.0, successful_trials = 996
gamma=0.33, alpa =0.33, epsilon=1.00: total_steps= 13139, net_rewards = 22456.5, negative_rewards = -183.5, successful_trials = 990
gamma=0.33, alpa =0.55, epsilon=0.10: total_steps= 12918, net_rewards = 22248.0, negative_rewards = -136.0, successful_trials = 995
gamma=0.33, alpa =0.55, epsilon=1.00: total_steps= 13283, net_rewards = 22506.0, negative_rewards = -157.5, successful_trials = 994
gamma=0.33, alpa =0.78, epsilon=0.10: total_steps= 12951, net_rewards = 22226.5, negative_rewards = -104.5, successful_trials = 994
gamma=0.33, alpa =0.78, epsilon=1.00: total_steps= 13070, net_rewards = 22369.0, negative_rewards = -150.5, successful_trials = 994

gamma=0.33, alpa =1.00, epsilon=0.10: total_steps= 12861, net_rewards = 22514.5, negative_rewards = -132.5, successful_trials = 992
gamma=0.33, alpa =1.00, epsilon=1.00: total_steps= 13447, net_rewards = 22347.0, negative_rewards = -118.0, successful_trials = 994
gamma=0.55, alpa =0.10, epsilon=0.10: total_steps= 13390, net_rewards = 22408.0, negative_rewards = -181.0, successful_trials = 991
gamma=0.55, alpa =0.10, epsilon=1.00: total_steps= 13990, net_rewards = 24129.0, negative_rewards = -817.0, successful_trials = 958
gamma=0.55, alpa =0.33, epsilon=0.10: total_steps= 13417, net_rewards = 22646.0, negative_rewards = -188.0, successful_trials = 989
gamma=0.55, alpa =0.33, epsilon=1.00: total_steps= 13697, net_rewards = 22390.0, negative_rewards = -206.5, successful_trials = 988
gamma=0.55, alpa =0.55, epsilon=0.10: total_steps= 13762, net_rewards = 22658.5, negative_rewards = -191.5, successful_trials = 991
gamma=0.55, alpa =0.55, epsilon=1.00: total_steps= 13437, net_rewards = 22384.5, negative_rewards = -204.0, successful_trials = 993
gamma=0.55, alpa =0.78, epsilon=0.10: total_steps= 12929, net_rewards = 22462.5, negative_rewards = -162.0, successful_trials = 992
gamma=0.55, alpa =0.78, epsilon=1.00: total_steps= 22734, net_rewards = 14650.5, negative_rewards = -232.5, successful_trials = 517
gamma=0.55, alpa =1.00, epsilon=0.10: total_steps= 12638, net_rewards = 22268.5, negative_rewards = -125.5, successful_trials = 992
gamma=0.55, alpa =1.00, epsilon=1.00: total_steps= 13713, net_rewards = 22364.5, negative_rewards = -158.5, successful_trials = 990
gamma=0.78, alpa =0.10, epsilon=0.10: total_steps= 13197, net_rewards = 23673.0, negative_rewards = -381.0, successful_trials = 984
gamma=0.78, alpa =0.10, epsilon=1.00: total_steps= 15081, net_rewards = 25157.5, negative_rewards = -1285.5, successful_trials = 925
gamma=0.78, alpa =0.33, epsilon=0.10: total_steps= 13278, net_rewards = 22822.5, negative_rewards = -233.5, successful_trials = 987
gamma=0.78, alpa =0.33, epsilon=1.00: total_steps= 13970, net_rewards = 23923.0, negative_rewards = -516.5, successful_trials = 977
gamma=0.78, alpa =0.55, epsilon=0.10: total_steps= 15838, net_rewards = 23080.0, negative_rewards = -500.5, successful_trials = 872
gamma=0.78, alpa =0.78, epsilon=0.10: total_steps= 13579, net_rewards = 22618.0, negative_rewards = -147.5, successful_trials = 991
gamma=0.78, alpa =0.78, epsilon=1.00: total_steps= 13445, net_rewards = 22625.0, negative_rewards = -207.0, successful_trials = 981
gamma=0.78, alpa =1.00, epsilon=0.10: total_steps= 13264, net_rewards = 22354.0, negative_rewards = -100.0, successful_trials = 996
gamma=0.78, alpa =1.00, epsilon=1.00: total_steps= 13133, net_rewards = 22388.0, negative_rewards = -127.5, successful_trials = 994
gamma=1.00, alpa =0.10, epsilon=0.10: total_steps= 13348, net_rewards = 22579.5, negative_rewards = -158.5, successful_trials = 993
gamma=1.00, alpa =0.10, epsilon=1.00: total_steps= 13942, net_rewards = 24432.5, negative_rewards = -1084.5, successful_trials = 953
gamma=1.00, alpa =0.33, epsilon=0.10: total_steps= 13132, net_rewards = 23790.5, negative_rewards = -399.0, successful_trials = 990
gamma=1.00, alpa =0.33, epsilon=1.00: total_steps= 13556, net_rewards = 22422.5, negative_rewards = -237.0, successful_trials = 984
gamma=1.00, alpa =0.55, epsilon=0.10: total_steps= 13299, net_rewards = 22530.0, negative_rewards = -186.0, successful_trials = 988
gamma=1.00, alpa =0.55, epsilon=1.00: total_steps= 13538, net_rewards = 22763.5, negative_rewards = -192.5, successful_trials = 995
gamma=1.00, alpa =0.78, epsilon=0.10: total_steps= 12945, net_rewards = 22523.0, negative_rewards = -122.5, successful_trials = 994
gamma=1.00, alpa =1.00, epsilon=0.10: total_steps= 13466, net_rewards = 22233.0, negative_rewards = -158.5, successful_trials = 990
gamma=1.00, alpa =1.00, epsilon=1.00: total_steps= 13416, net_rewards = 22235.0, negative_rewards = -145.0, successful_trials = 996

As per Wikipedia: Learning factor (alpa) “A factor of 0 will make the agent not learn anything, while a factor of 1 would make the agent consider only the most recent information. In fully deterministic environments, a learning rate of 1 is optimal.”

Discount Factor(gamma): “The discount factor γ\gamma determines the importance of future rewards. A factor of 0 will make the agent "myopic" (or short-sighted) by only considering current rewards, while a factor approaching 1 will make it strive for a long-term high reward.”

Alpa of 0.78, Gamma of 0.1 and Epsilon of 0.1 gave the best possible performance, the best run has successfully trials of 998 out of 100 with negative rewards of just 66, with net total rewards of 22346.

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?

I would define optimal policy by measuring the number of steps, total net rewards, total negative rewards, total successfully completed trials that it achieves as it

moves along to the destination. Lower number of steps and higher rewards indicate optimal policy.
The Q-Learning method receives lower total negative rewards as it learns the reward system while moving to the destination. The following graph shows that smartcab initially accumulates negative rewards rapidly(steep slope) and as it learns the environment, rate of negative rewards decreases(flat slope) as seen in the graph below.
x-axis: number of steps taken so far
y-axis: negative rewards