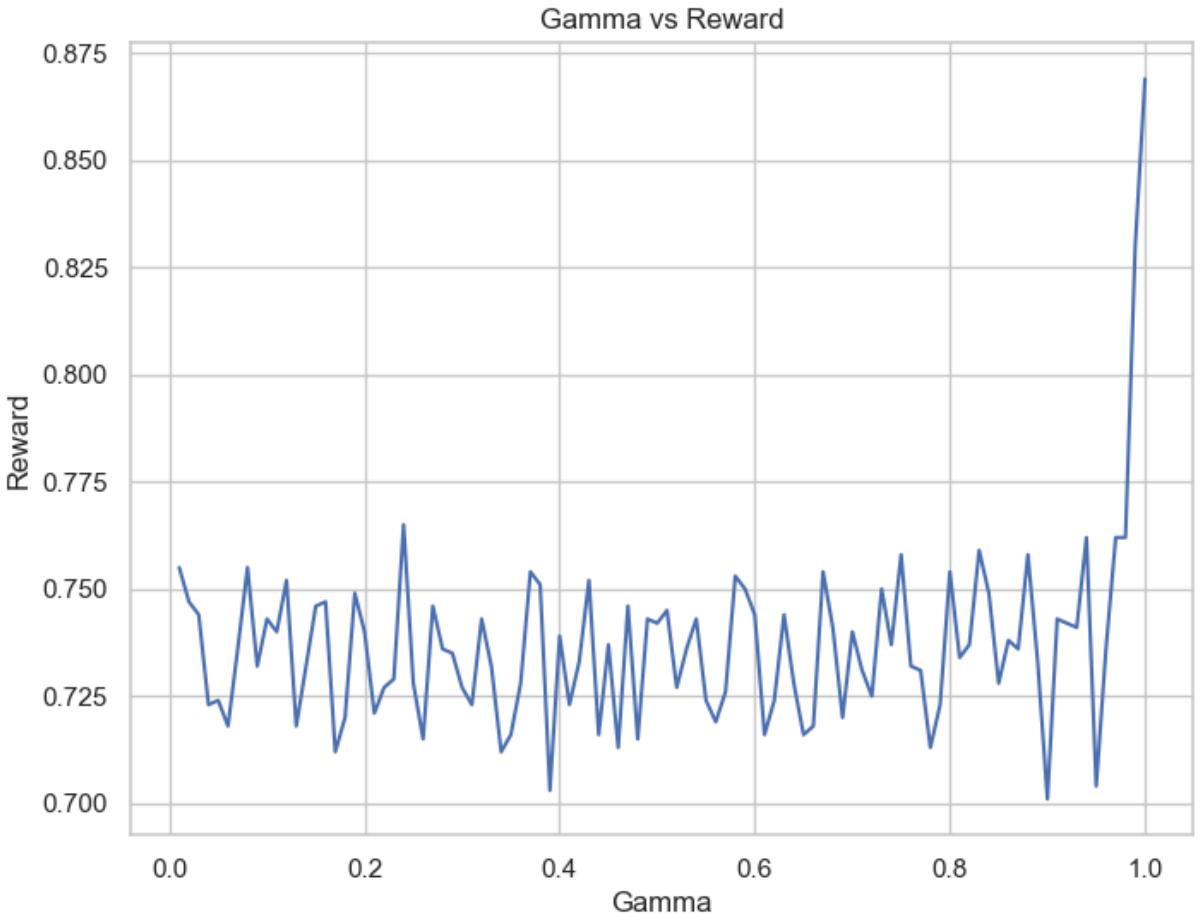
Задание 3

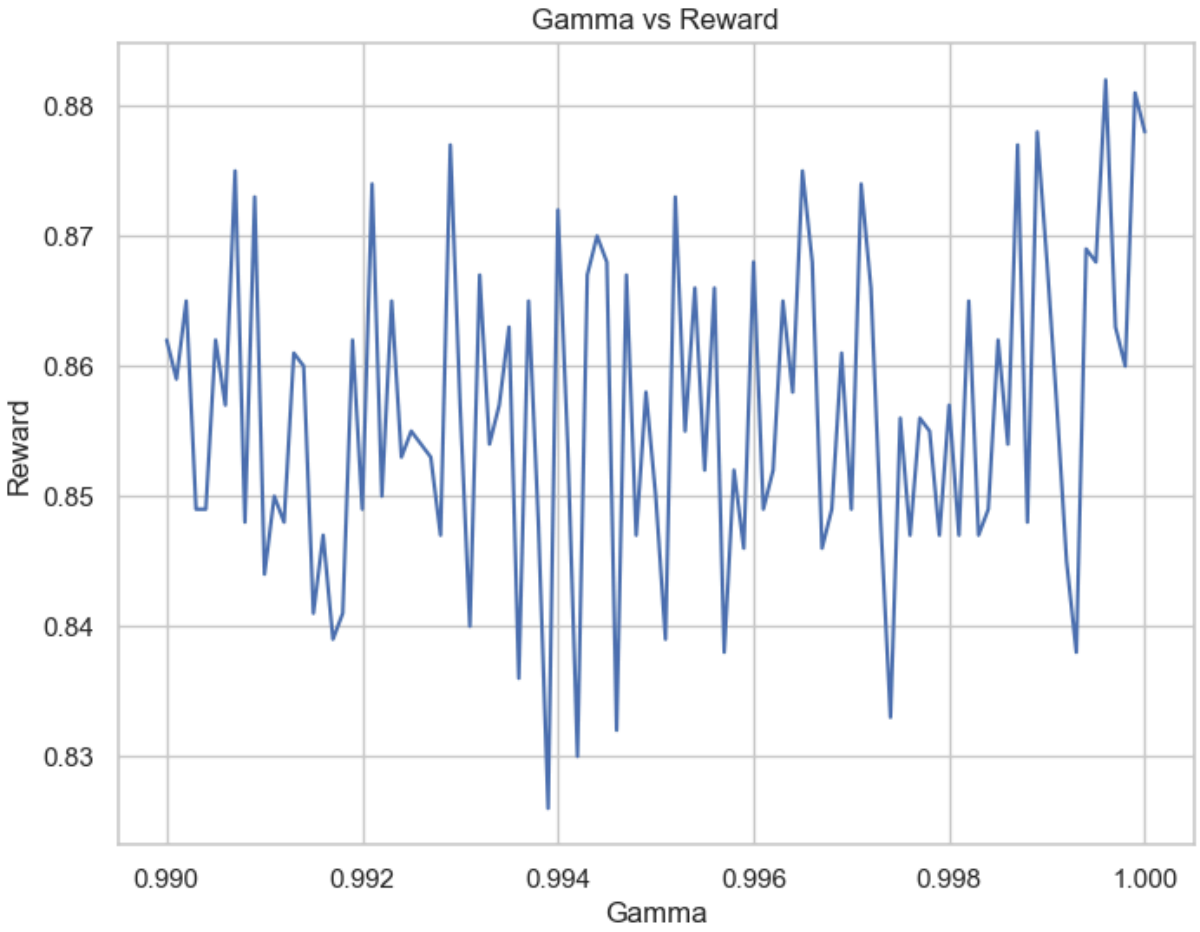
В алгоритме Policy Iteration важным гиперпараметром является gamma. Требуется ответить на вопрос, какой gamma лучше выбирать. Качество обученной политики можно оценивать например запуская среду 1000 раз и взяв после этого средний total_reward.

In [28]:



Проанализируем все значения gamma от 0.01 до 1 с шагом 0.01, чтобы понять в каком диапазоне искать оптимальное значение gamma. График показывает что почти все значения выдают одинаковое (с точностью до шума) значение средней награды, в районе 0.75. Исключение составляют значение близкие к единице, включая саму единицу, которые дают среднюю награду около 0.85. Проанализируем большее точно интервал от 0.99 до 1.

In [30]:



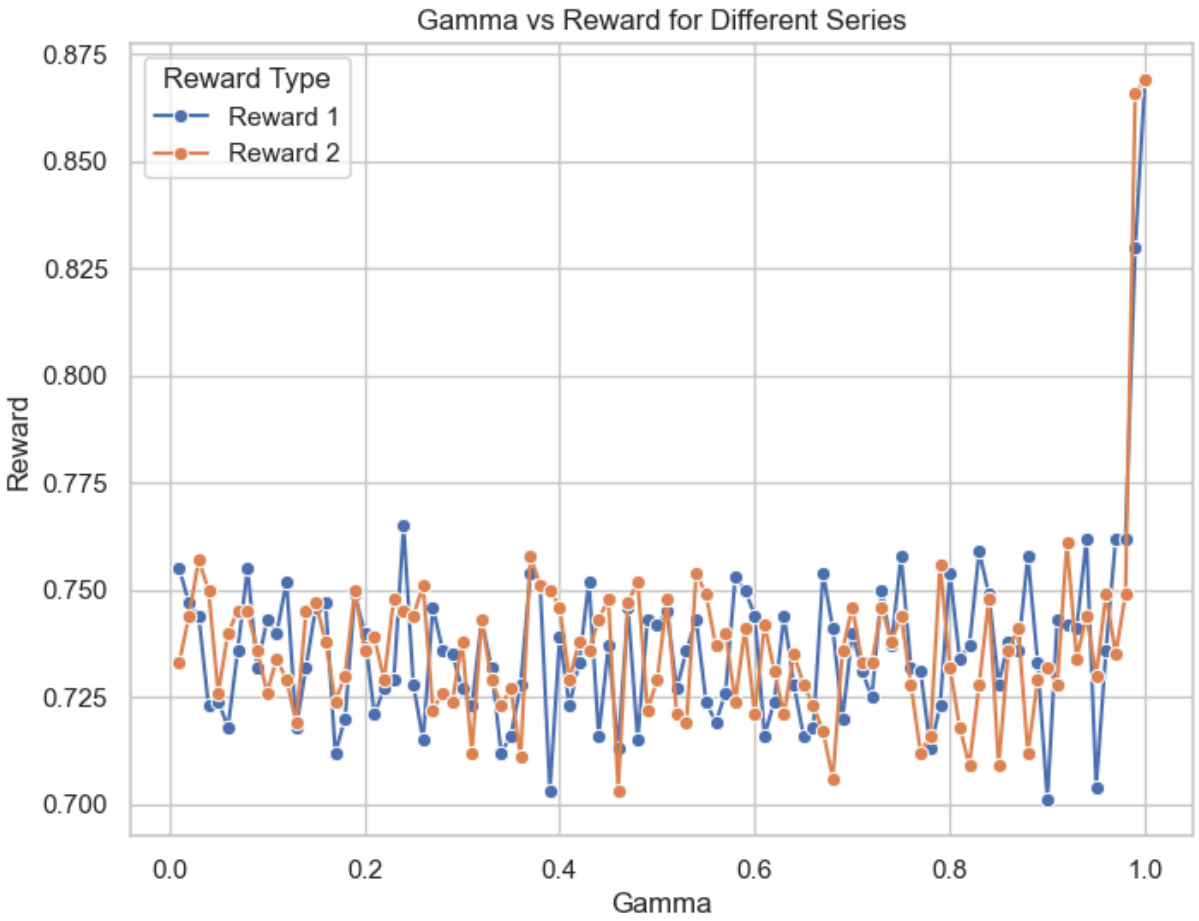
На этом графике видно что почти все значения дают примерно один и тот же результат.

Результат:

- Оптимальное значение gamma = 1.
- Поскольку теоретические гарантии есть только для gamma меньше единицы, можно использовать значения очень близкие к 1, например, 0.9999, но данная среда достаточно простая и, кажется, 1 вполне остаточно.

На шаге Policy Evaluation мы каждый раз начинаем с нулевых values. А что будет если вместо этого начинать с values обученных на предыдущем шаге? Будет ли алгоритм работать? Если да, то будет ли он работать лучше?

In [33]:



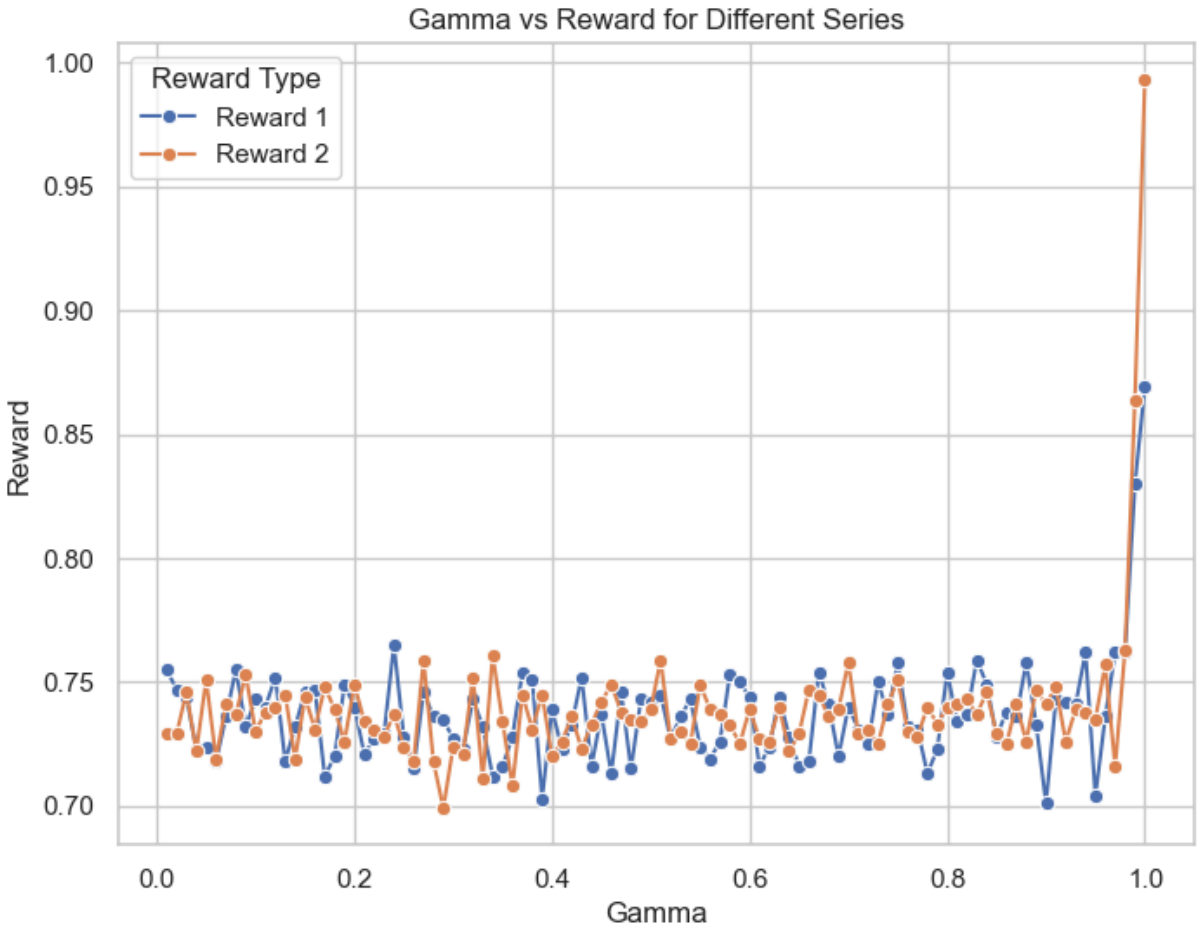
На графике не видно существенных различий в обучении алгоритма для любого значения gamma. Алгоритм, основанный на теореме о неподвижной точке для сжимающего отображения гарантирует достижение неподвижной точки из любых значений, поэтому какими бы мы не выбрали значения value, результат будет одинаковый.

Результат:

- Алгоритм работает, если переиспользовать обученные значения value.
- Алгоритм работает так же эффективно.

Написать Value Iteriation. Исследовать гиперпараметры (в том числе gamma). Сравнить с Policy Iteration. Поскольку в Policy Iteration есть еще внутренний цикл, то адекватным сравнением алгоритмов будет не графики их результативности относительно внешнего цикла, а графики относительно, например, количества обращения к среде.

In [36]:



Value iteration показывает почти такие же результаты как и policy iteration для всех значений gamma кроме 1. Кажется, что разброс значений у value iteration меньше, то есть алгоритм стабильнее, хотя это трудно измерить. Для значения gamma = 1 value iteration показал существенно лучший результат, более 0.99. Policy iteration в данном примере имел 20 шагов на policy evaluation и 20 эпох, за это время, очевидно, не удалось найти безопасную стратегию которая не будет делать потенциальные шаги в пропасть, в то время как value iteration имел 400 шагов на обновление value function, и успел найти верные значения, чтоб избегать опасных маневров. Сложно сравнивать подобные алгоритмы, так как они находятся не в равных условиях, при достаточном числе шагов они оба сойдутся к оптимальной стратегии, но для данной простой среды value iteration делает это быстрее.

Результат:

- value iteration показывает лучший результат, чем policy iteration для данной среды при gamma = 1.