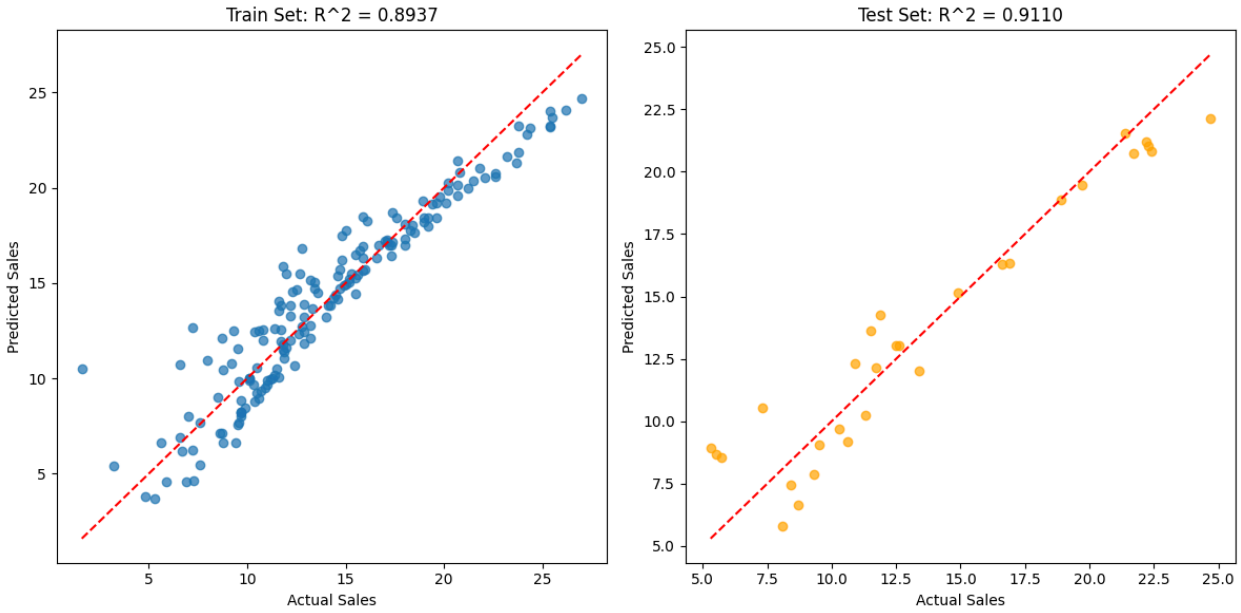
1. The optimized weights of my gradient descent are:
   $\theta_0 = 14.02477267$, $\theta_1 = 3.83763432$, $\theta_2 = 2.79115525$, $\theta_3 = 0.01635503$

   $$h\theta(x) = 14.02477267 + 3.83763432 \cdot TV + 2.79115525 \cdot Radio + 0.01635503 \cdot Newspaper$$

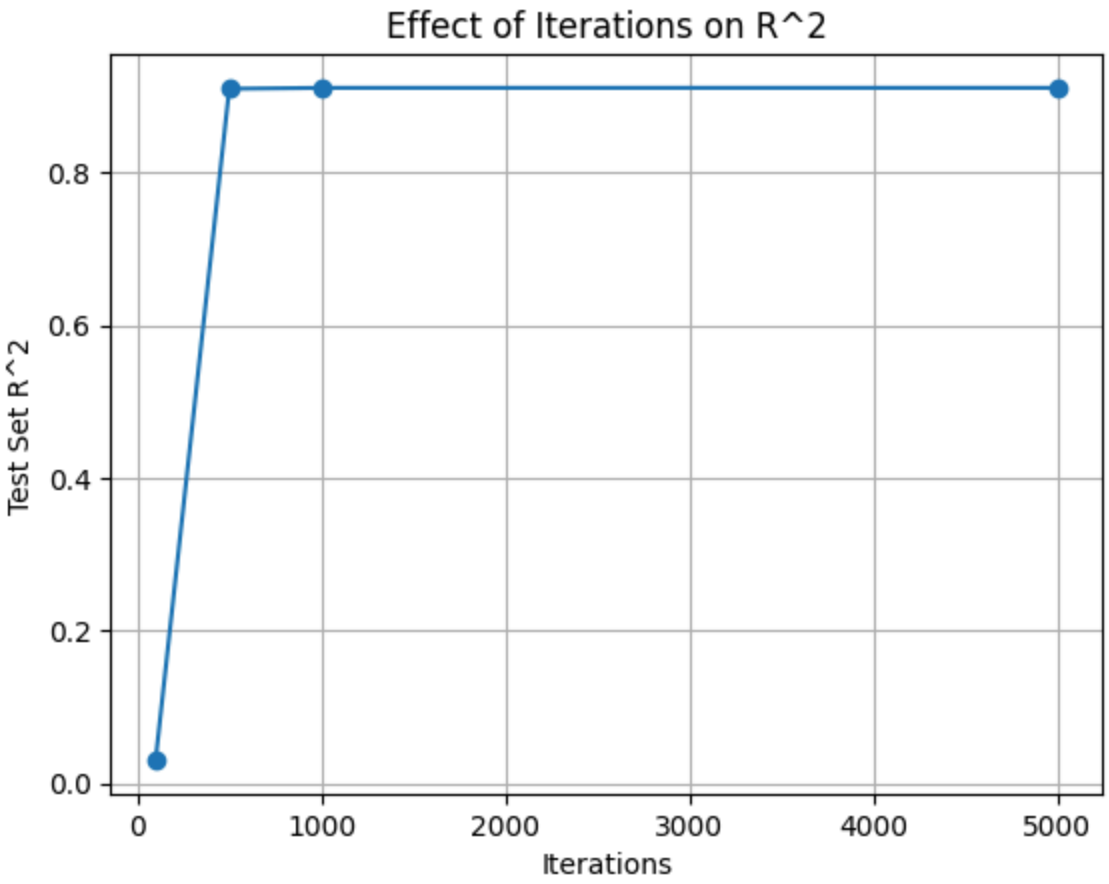
   This shows that the predictor TV has 3.84 influence on the target Sales. In other words, advertising through TV contributes the most to a company's sales based on the advertising dataset. TV is followed by Radio, and the medium that affects the sales the least is the Newspaper.

2. Provide a scatter plot of the hat($y^i$) and $y^i$ for both the train and test set. Is there a trend? Provide an $r^2$ score (also available in sklearn).



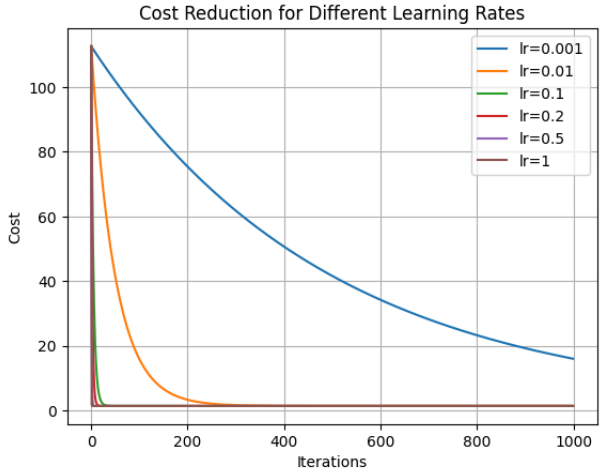

   The graph shows the $R^2$ scorers for training and test sets. As we can see, there is a small difference between both scores wish indicates that the model performs well in both training and testing sets. Both training and testing sets also follows a linear trend which explains the relationship of the predictors and the target feature.

3. What happens to the error, r2, and cost as the number of iterations increase? Show your data and proof. You can alternatively plot your result data for visualization and check until 50000 iterations or more (actually).




The graph shows $R^2$ as the number of iterations increase up to 5000. This shows that $R^2$ significantly increases from 0 to 0.9110 within just 1000 iterations. The flat line above indicates that the model already reached its global minimum or convergence around those points.

4. Once you determine the optimal number of iterations, check the effect on the cost and error as you change the learning rate. The common learning rates in machine learning include 0.1, 0.01, 0.001, 0.0001, 0.2 but you have the option to include others. Visualize the cost function (vs the optimal number of iterations) of each learning rate in ONLY ONE PLOT. Provide your analysis.



This implementation shows how learning rates affect the cost across the iteration. This implies that the bigger the learrning rate, the steeper the cost function across the iteration. Additionaly, upon experimenting I found out that the model started to reach its global minimum somewhere between 0.01 to 0.1 learning rate with 1000 iterations.

5. Is there a relationship on the learning rate and the number of iterations?

   The value of the learning rate is inversely proportional to the number of iterations. This means that the higher the learning rate, the lower the number of iterations. In other words, the model can reach the convergence with higher learning rate but might skip convergence at some point (overshooting).

6. Compare the results with the results of ordinary least squares function.

| **Weights** | $\theta_0$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $R^2$ |
|---|---|---|---|---|---|
| **Gradient Descent** | 14.02477267 | 3.83763432 | 2.79115525 | 0.01635503 | 0.911027570 2091712 |
| **Ordinary Least Squares** | 14.02477267 | 3.83763432 | 2.79115525 | 0.01635503 | 0.911027570 2091711 |

As we can see, both Gradient Descent and OLS achieved similar optimal weights for all the predictors. This means that one can be used as an alternative to minimize the cost function for linear regression.