**Activity 1: Normalization**

**Answer – How does normalization affect the feature values?**  
Normalization transforms the features to have a mean of 0 and a standard deviation of 1. This ensures all features contribute equally during training, especially in gradient descent. It removes scale differences and speeds up convergence. For example, hours ranging from 0–10 become values like -1.23, 0.01, 1.56, making learning more stable and efficient.

**Activity 2: MSE vs MAE**

**Answer – Why does MSE penalize larger errors more than MAE?**  
MSE (Mean Squared Error) squares the error term, so larger errors are amplified exponentially. For instance, an error of 2 contributes 4 to MSE but only 2 to MAE (Mean Absolute Error). This makes MSE more sensitive to outliers, while MAE treats all errors linearly and equally. MSE is preferred when large errors are especially undesirable.

**Activity 3: Gradient Descent**

**Answer – How does the learning rate affect convergence?**  
The learning rate controls how big a step is taken toward the optimal solution in each iteration. A small learning rate (e.g., 0.001) results in slow convergence and might get stuck in local minima. A large one (e.g., 0.1) might overshoot the minimum and diverge. An optimal rate (e.g., 0.01) balances speed and stability.

**Activity 4: Real vs. Synthetic Data**

**Answer – Why might the model perform differently on real vs. synthetic data?**  
Synthetic data follows a clean linear pattern (e.g., exam score = 2×hours + noise), so it's easier to fit. Real-world data, like housing prices, has noise, multicollinearity, outliers, and non-linear patterns. Thus, the model's performance (MSE/MAE) may degrade due to the complexity and messiness of real-world data.

QUESTIONS

**1. How does the choice of cost function (MSE vs. MAE) affect optimization?**

* MSE is smooth and differentiable everywhere, making it more suitable for gradient-based optimization.
* MAE’s derivative is constant (±1) and undefined at zero, making optimization slower or requiring special handling.
* MSE leads to faster convergence but is sensitive to outliers. MAE is more robust to outliers but slower to optimize.

**2. What challenges arise when scaling to multiple features?**

* Feature scaling becomes crucial to ensure no feature dominates.
* Computational cost increases.
* Gradient descent might converge slowly or diverge without proper learning rate adjustment.
* Data with multicollinearity can confuse the model.

**3. How does gradient descent compare to scikit-learn’s built-in linear regression?**

* sklearn.LinearRegression uses the Normal Equation (closed-form) which is faster and more accurate for small datasets.
* Gradient descent is more flexible, scales better for large datasets, and is needed when the closed-form is computationally expensive.
* Gradient descent allows for customization (e.g., learning rate, regularization, batching).