Differentiate between logistic regression and linear regression through two real-world scenarios.

Hint: Differentiate in terms of i) Definition, ii) Datasets compatibility, iii) Model, iv) Validation Metrics, v) Visualization ( through graphs)

Both **logistic regression** and **linear regression** are widely used statistical techniques in machine learning, but they are used for different types of prediction tasks. Here's a detailed comparison using two real-world scenario

Scenario 1: Predicting House Prices (Linear Regression)

Let's consider a scenario where you want to predict house prices based on various features like square footage, number of bedrooms, etc.

1. i) Definition:

Linear Regression: Linear regression is used when the target variable (dependent variable) is continuous and numerical. It finds the best-fit line to predict values for the target.

Logistic Regression: Logistic regression is used when the target variable is categorical, typically for binary classification tasks.

1. ii) Datasets Compatibility:

Linear Regression: Works with continuous data (e.g., predicting house prices, temperature, etc.). The dependent variable is numeric.

* + - Example Dataset: House features such as square footage, number of bedrooms, etc., are used to predict house prices (continuous value).

Logistic Regression: Works with categorical data (e.g., classification problems like spam detection, medical diagnoses). The dependent variable is binary (0 or 1).

* + - Example Dataset: Whether a house will be sold (yes/no), or if a patient has a disease (0 = No, 1 = Yes).

1. iii) Model:

Linear Regression: The model finds a linear relationship between the independent variables and the dependent variable, using a line of best fit. The equation is of the form:  
y=β0+β1x1+β2x2+⋯+βnxny = \beta\_0 + \beta\_1x\_1 + \beta\_2x\_2 + \dots + \beta\_nx\_ny=β0​+β1​x1​+β2​x2​+⋯+βn​xn​

Logistic Regression: The model predicts probabilities (between 0 and 1) for classification problems, using a logistic (sigmoid) function. The equation is: P(y=1)=11+e−(β0+β1x1+⋯+βnxn)P(y=1) = \frac{1}{1 + e^{-(\beta\_0 + \beta\_1x\_1 + \dots + \beta\_nx\_n)}}P(y=1)=1+e−(β0​+β1​x1​+⋯+βn​xn​)1​

1. iv) Validation Metrics:

Linear Regression:

Common validation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R²).

R² shows the proportion of the variance in the dependent variable that is predictable from the independent variables.

Logistic Regression:

Common validation metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix, AUC-ROC.

These metrics help evaluate the classification model’s performance on how well it predicts binary outcomes.

v) Visualization:

Linear Regression: The best-fit line can be visualized on a scatter plot, where the dependent variable is continuous, and the regression line shows the predicted values.

Graph: A straight line plotted against data points.

Logistic Regression: Visualization involves plotting a sigmoid curve that represents the probability of the binary outcome.

Graph: A curve (sigmoid curve) showing the probability of classifying data into a category.

Scenario 2: Predicting Whether an Email is Spam or Not (Logistic Regression)

In this scenario, the goal is to predict whether an email is spam (1) or not spam (0) based on features like the presence of certain words, sender's address, etc.

1. Definition:

Linear Regression: Again, linear regression is used when the target is continuous. Here, linear regression would be inappropriate because the target is binary (spam or not).

Regression: Logistic regression is ideal for binary classification tasks, where the target is categorical (yes/no, 1/0).

1. ii) Datasets Compatibility:

Linear Regression: Would be used if you were predicting a continuous value, such as predicting the probability of an email being spam (not suitable for binary outcomes directly).

Logistic Regression: Perfectly suited for datasets with binary outcomes, such as predicting whether an email is spam (1) or not spam (0).

1. iii) Model:

Linear Regression: For this task, linear regression would predict a continuous score, not a binary outcome. It would give a value that’s not constrained between 0 and 1, which is not ideal for classification.

Logistic Regression: Logistic regression will output a probability between 0 and 1 (after applying the sigmoid function), which can be thresholded to classify the email as spam (1) or not (0).

1. iv) Validation Metrics:

Linear Regression: Would be inappropriate here since we're working with a classification problem.

Logistic Regression:

Key metrics would be: Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

Accuracy measures the overall percentage of correctly classified emails.

Precision and Recall help evaluate how well the model distinguishes between spam and non-spam.

Visualization:

Linear Regression: A straight line wouldn’t be an effective visualization in this case, as it doesn’t handle probabilities well for binary classification.

Logistic Regression: You would visualize the sigmoid function, where the x-axis represents the input features (e.g., frequency of certain words), and the y-axis shows the probability of the email being spam.

Graph: A sigmoid curve that shows the likelihood of spam (values between 0 and 1).