DS Tutorial-1

Question 1:
Based on advertising data, find out the residual standard error(RSE), R^2 and F-statistics with respect to TV, radio, newspaper advertising budgets. Comment on the values.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Answer:
Residual Standard Error = 3.26: This value represents the average deviation of the actual data points from the regression line. In other words, it indicates the typical size of the errors made by the model in predicting the dependent variable (sales) based on the advertising budgets. A lower value for the residual standard error is generally desirable, as it suggests that the model fits the data more closely.
R-squared = 0.612: This statistic, expressed as a percentage (61.2%), tells us the proportion of the variance in the dependent variable (sales) that is explained by the independent variables (TV, radio, newspaper advertising budgets) in the regression model. In this case, 61.2% of the variation in sales can be attributed to the combined influence of these advertising channels. A higher R-squared value indicates a better fit of the model to the data.
F-statistic = 312.1: This statistic tests the overall significance of the regression model. It compares the variance explained by the model to the residual variance. A high F-statistic value (as in this case) suggests that the regression model is statistically significant, meaning that the independent variables (advertising budgets) collectively have a significant impact on the dependent variable (sales).

Question 2:
Create a dataset of your own choice, explain the dataset and using logistic regression predict the value for unknown inputs.

Answer:
Let's create a dataset to predict whether a student will pass or fail an exam based on their study hours.
Data:

Study Hours (X)	Pass (Y)
2	0

4	0
6	1
8	1
10	1
12	1
14	1
16	1

X: Study Hours (independent variable)

Y: Pass (dependent variable, binary: 0 = Fail, 1 = Pass)

Logistic Regression:

Logistic regression is suitable for this binary classification problem. We'll model the probability of passing **(P(Y=1))** as a function of study hours using the sigmoid function:

P(Y=1) = 1 / (1 + exp(-(b0 + b1*X)))

where:

- **b0:** Intercept
- **b1:** Coefficient for study hours

To predict whether a student with unknown study hours will pass, we plug the study hours value into the fitted model and calculate the probability. If the probability is above a certain threshold (e.g., 0.5), we predict that the student will pass; otherwise, we predict that they will fail.

Implementation:

Python

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
X = np.array([[2], [4], [6], [8], [10], [12], [14], [16]])
y = np.array([0, 0, 1, 1, 1, 1, 1, 1])
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
intercept = model.intercept_
coef = model.coef_
```

BY
KEERTHANA SURESH.D
S5 IT
ROLL NUMBER:33