Data Science 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.

Answer:
Residual Standard Error (RSE): 3.26
This indicates the average deviation of the observed data points from the regression line. A lower RSE implies a better fit. However, its interpretation depends on the scale of the dependent variable.
R^2 (Coefficient of Determination): 0.612
This means that 61.2% of the variance in the dependent variable (e.g., sales) is explained by the independent variables (TV, radio, and newspaper advertising budgets). While this is a reasonable value, it suggests that about 38.8% of the variance is unexplained, possibly due to other factors not included in the model.
F-statistic: 312.1
This statistic tests whether at least one predictor has a non-zero coefficient. A high F-statistic with a significant p-value indicates that the model as a whole is statistically significant. Given the large value of 312.1, the model appears to have strong predictive power.

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

Answer:
Dataset where a company wants to predict if a customer will purchase a product based on their **age**, **annual income**, and **spending score**.

Dataset Explanation

Age: The age of the customer (18–65 years).
Annual Income: Income in thousands of dollars.
Spending Score: A score (1–100) that reflects customer spending behavior.
Purchased: Binary target variable (1 = Purchased, 0 = Not Purchased).

Create the dataset using Python
Generate the data and build a logistic regression model to predict whether a customer will purchase the product based on the input features.

```
# New inputs for prediction
new_inputs = pd.DataFrame({
    'Age': [30, 50],
    'Annual_Income': [85, 40],
    'Spending_Score': [70, 30]
})

# Predict using the logistic regression model
```

predictions = log_reg.predict(new_inputs)

Here is a sample of the generated dataset:

Age	Annual_Income	Spending_Score	Purchased
56	27	62	0
46	107	58	1
32	82	52	1
60	30	12	0
25	100	39	1

Age: Customer age ranges from 18 to 65.
Annual Income: Income ranges from $20k to $120k.
Spending Score: Spending score ranges from 1 to 100.
Purchased: Binary outcome where 1 indicates a purchase.

After training the logistic regression model:
Accuracy: 90% on the test set.
Classification Report:

Metric	Class 0	Class 1
Precision	0.86	1.00
Recall	1.00	0.73
F1-Score	0.93	0.84

Overall: The model performs well, especially for customers who did not purchase (Class 0). However, it is slightly less effective at identifying customers who made a purchase (Class 1)

Predictions for New Inputs:

1. Age = 30, Annual Income = 85, Spending Score = 70
   Predicted Outcome: Purchased (1)
   This customer is likely to purchase the product due to their high spending score and income.
2. Age = 50, Annual Income = 40, Spending Score = 30
   Predicted Outcome: Not Purchased (0)
   This customer is less likely to purchase the product due to their lower spending score and income.