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.

Answer:

	TV
Beta1	0.04753664
Beta0	7.03259355
RSS	2102.53856
RSE	3.26
TSS	5417.14875
R^2	0.61187358
F-statistic	312.143059

	Radio
Beta1	0.202495783
Beta0	9.311638095
RSS	3618.479549
RSE	4.27
TSS	5417.14875
R^2	0.332032455
F-statistic	98.42158757

	Newspaper
Beta1	0.054693098
Beta0	12.35140707
RSS	5134.804544
RSE	5.09
TSS	5417.14875
R^2	0.052120445
F-statistic	10.88729908

Conclusion:
TV: The strongest predictor, with the lowest RSE (3.26), high R^2 (0.6119), and a very significant F-statistic (312.14).
Radio: A moderate predictor, with a higher RSE (4.27), moderate R^2 (0.3320), and a significant F-statistic (98.42).
Newspaper: The weakest predictor, with the highest RSE (5.09), very low R^2

(0.0521), and the lowest F-statistic (10.89).

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.