Easy: Analyzing Sales Data

Question:
You are provided with a dataset containing sales records of a retail store over the past year. The dataset includes columns for date, product category, product ID, sales amount, and location. Your task is to perform basic exploratory data analysis (EDA) to answer the following questions:

1. What was the total sales amount for each month?
2. Which product category generated the highest revenue?
3. Identify the top 5 products by sales amount.
4. Visualize the sales trends over the year.

Dataset:

- Date: The date of the sale
- Product_Category: The category of the product
- Product_ID: The unique identifier of the product
- Sales_Amount: The amount of sales in dollars
- Location: The location of the store

Expected Output:

- Monthly sales totals
- Bar chart of revenue by product category
- List of top 5 products by sales amount
- Line graph showing sales trends over the year

Medium: Predicting House Prices

Question:
Using a dataset of housing prices, develop a model to predict the price of a house based on its features. The dataset includes various attributes of the houses such as square footage, number of bedrooms, number of bathrooms, location, and more. Your task is to:

1. Clean and preprocess the data (handle missing values, encode categorical variables, etc.).
2. Perform feature engineering if necessary.
3. Train and evaluate multiple regression models (e.g., linear regression, decision tree regression, etc.).
4. Select the best-performing model and justify your choice.

Dataset:

- Square_Footage: The size of the house in square feet
- Bedrooms: Number of bedrooms
- Bathrooms: Number of bathrooms
- Location: The location of the house (categorical)
- Year_Built: The year the house was built
- Price: The price of the house

Expected Output:

- Preprocessed dataset
- Comparison of model performance metrics (e.g., RMSE, MAE)
- Final selected model with justification

Difficult: Sentiment Analysis on Social Media Data

Question:
You are given a dataset of tweets containing customer feedback about a product. Your task is to build a sentiment analysis model to classify the tweets into positive, negative, and neutral sentiments. This involves:

1. Data cleaning and preprocessing (removing stop words, tokenization, etc.).
2. Exploratory data analysis to understand the distribution of sentiments.
3. Feature extraction using techniques such as TF-IDF or word embeddings.
4. Training and evaluating multiple machine learning models (e.g., logistic regression, Random Forest, neural networks).
5. Fine-tuning the best-performing model and evaluating its performance on a test set.

Dataset:

- Tweet_ID: The unique identifier of the tweet
- Tweet_Text: The text content of the tweet
- Sentiment: The sentiment label (positive, negative, neutral) [provided for training]

Expected Output:

- Cleaned and preprocessed text data
- EDA reports and visualizations showing sentiment distribution
- Comparison of model performance (e.g., accuracy, F1-score)
- Final sentiment analysis model with evaluation on test data