Problem Statement

Background & Context

This assignment explores how probability distributions and hypothesis testing are applied to real-world industrial scenarios using data-driven approaches. By simulating random events and analyzing actual used car sales data (car_sales_cleaned.csv), we aim to understand patterns in user behavior, product performance, and market trends. These statistical techniques are widely used across industries—for instance, in forecasting user sign-ups, tracking manufacturing errors, and validating business assumptions like pricing strategies.

Objective

The objective is to build practical skills in simulating and interpreting key probability distributions (Binomial, Poisson, Chi-Square) and applying hypothesis tests (One-Way ANOVA, Two-Way ANOVA, Chi-Square Test) to real datasets. This helps learners make data-backed decisions and draw actionable insights in contexts like quality control, marketing analysis, and automotive resale.

Part 1: Probability Distributions

Tasks:

1. **Binomial distribution** A coin is tossed 5 times. Simulate the experiment in Python using the function np.random.binomial() to find the distribution of the Number of times heads come. Run the experiment 10, 100 and 1000 times and compare it with the theoretical binomial distribution for the same case.

2. **Poisson Distribution distribution** A website receives an average of 3 user sign-ups per hour. Simulate this process in Python using the np.random.poisson() function to find the distribution of the number of sign-ups per hour. Run the experiment 10, 100, and 1000 times and compare it with the theoretical Poisson distribution for the same mean (λ = 3).

3. **Chi-Square Distribution**Suppose a factory tracks small measurement errors in a machine's output. These squared errors are known to follow a Chi-Square distribution with 2 degrees of freedom. Simulate this process in Python using np.random.chisquare() to model the distribution of these squared errors. Run the simulation with 10, 100, and 1000 measurements. Plot the histograms and compare them with the theoretical Chi-Square PDF.

Part 2: Hypothesis Testing

Tasks:

1. **One-Way ANOVA:**

   A popular belief in the automotive resale industry is that a **Toyota** car typically resells for **₹12,400** on average in the used car market. You have access to a detailed and cleaned data set of used car sales stored in the file car_sales_cleaned.csv.

   Using the data provided, test whether this belief about the **average resale price of Toyota cars** holds true. (**Use an appropriate hypothesis test and a significance level of 0.05 to support your conclusion.)**

2. **Two-Way Anova:**

   A used car company wants to know what factors influence the resale price of a black color car. They believe that two things might affect the price:

   - The body type of the car (like SUV, Sedan, Coupe, etc.)
   - The transmission type (Automatic or Manual)

   They also want to know if the combination of body type and transmission has any extra effect.

   (Use a Two-Way Anova significance level of **0.05**. Take a sample of 5000 cars.

3. **Chi-Square Test:**

   Is the distribution of Toyota car colors the same as that of all cars? Using the same data, test whether there is a statistically significant difference between the distribution of car colors for Toyota vehicles and the overall market.

   Use a Chi-Square Test of Independence with a significance level of 0.05