Final Assignment: Data Cleaning & Analytics

Context:
You are a **Data Analyst** at a retail company. The dataset contains demographic, financial, and behavioral information about customers. Your job is to **clean the data and extract meaningful insights** to help the business improve decision-making.

Part 1: Exploratory Data Analysis (EDA)

1. Load the dataset and display the first 10 rows.
2. Show dataset shape (#rows, #columns).
3. Generate summary statistics for numerical features.
4. Count missing values per column and calculate their percentage.
5. Identify categorical features and list unique values.
6. Check for duplicate records.

Part 2: Handling Missing & Inappropriate Data

1. Identify and impute missing values:
    - Numerical → median/mean.
    - Categorical → mode or "Unknown".
2. Find invalid ages (<10 or >100) and treat them as missing.
3. Correct invalid incomes (negative or >1,000,000).
4. Ensure purchase counts are non-negative integers.
5. Validate gender and city columns for unexpected categories.

Part 3: Handling Outliers

1. Detect outliers in `AnnualIncome` and `SpendingScore` using boxplots.
2. For **CustomerSatisfactionScore**, decide whether it is closer to a normal distribution and handle outliers (e.g., Z-score method).
3. For **LastPurchaseAmount**, decide whether it is skewed and handle outliers (e.g., IQR/ Tukey's method).
4. Justify why you used different methods.

Part 4: Insights & Analytics

1. Find the top 5 cities by **average spending score**.
2. Compare **average annual income** across cities.
3. Analyze the correlation between **AnnualIncome** and **SpendingScore** (before and after handling outliers).
4. Compare **average purchase count by gender**.
5. Which **age group** (Young <30, Middle 30–55, Senior >55) has the highest spending score?
6. Find top 5 customers with the **highest LastPurchaseAmount**.
7. Compare **CustomerSatisfactionScore** across cities – which city has the most satisfied customers?
8. Do customers with high satisfaction scores also spend more on average?
9. Find the relationship between **PurchaseCount** and **CustomerSatisfactionScore**.
10. Identify which gender shows the **highest repeat purchases (PurchaseCount)**.