**Case Study: SmartMart’s Data Challenge**

# Background

SmartMart is a mid-sized retail chain with stores across Tier 1 and Tier 2 cities. The company has been collecting transactional and loyalty program data for over a year. Now, management wants to use AI and ML to **predict sales and customer churn** — but the dataset they pulled from their systems is far from ready.

## Issues Noticed:

* Some data are missing or unrealistic.
* Data has inconsistent formats
* Data names have typos and inconsistent spellings.
* Outliers exist
* Several predictions seem suspiciously high or negative.

## The dataset contains the following fields:

|  |  |
| --- | --- |
| **Customer\_ID** | Unique customer identifier (but duplicates might exist). |
| **Age** | Age in years. |
| **Annual\_Income** | Annual household income in local currency. |
| **Spending\_Score** | Internal score (0–100) for customer engagement. |
| **Loyalty\_Program** | Whether the customer is in the loyalty program. |
| **City** | City of residence or primary shopping location. |
| **Total\_Purchases** | Number of purchases in the past year. |
| **Last\_Month\_Spend** | Amount spent last month. |
| **Monthly\_Sales\_Prediction** | Predicted next-month spend from an internal model. |
| **Will\_Churn** | Yes/No if the customer is expected to stop shopping. |

# Tasks at hand

## Step 1 – Preprocessing

* Identify and handle **missing values** in all columns.
* Standardize formats.
* Remove or correct unrealistic values.
* Treat outliers where necessary.
* Verify and remove duplicate records.

## Step 2 – Exploratory Data Analysis (EDA)

* Summarize numerical variables (mean, median, quartiles).
* Explore correlations between Annual\_Income, Spending\_Score, and Total\_Purchases.
* Compare churn vs non-churn groups on key metrics.
* Identify top cities by total spend.

## Step 3 – Insights & Recommendations

Use your findings to provide management with **3–5 actionable recommendations** for improving sales and reducing churn.

### Additional Questions

1. After cleaning, what is the **average annual income** and **average spending score** of SmartMart customers?
2. Which **3 cities** have the highest **total purchases**?
3. Is there a significant difference in Annual\_Income between churned and non-churned customers?
4. What is the correlation between Spending\_Score and Monthly\_Sales\_Prediction?
5. Which customers are most likely to churn, and why?
6. Identify **outliers** in Last\_Month\_Spend. How do they impact the overall average?
7. Do higher incomes always lead to higher spending? Support your answer with analysis.
8. Does the existing Monthly\_Sales\_Prediction seem reliable? How would you improve it?
9. Which single factor (or combination) seems to influence churn the most?
10. Based on your analysis, what **three strategic actions** would you recommend to SmartMart?