Sales Data Analysis - Python (class project)

Executive Summary

Project: Sales Data Analysis + Visualization Dashboard (OMIS 114 Capstone)
Team: Charlie G, Alex S, Lucas I

Objective:

Turn a raw sales dataset into actionable insights + a modeling pipeline to support product focus and retention decisions.

Data:

- 10,000 rows × 15 columns (numeric, categorical, and time fields)
- Engineered features: **customer_lifetime_value** and **purchase_recency**

What we built:

- Cleaning pipeline: missing values + outliers (IQR capping + targeted median replacements)
- EDA + visual insights (age distribution, revenue by category, brand/location analysis)
- ML: Random Forest churn classifier (churn = 180+ days no purchases) + regression for customer value

Key results:

- Churn model performance: ROC AUC ≈ **0.79**
- Category concentration: **Automotive + Electronics ≈ 52% of revenue** → focus strategy

Business recommendations:

- Use churn predictions to trigger targeted retention offers for at-risk customers
- Reduce low-performing categories (e.g., books/beauty) and streamline suppliers toward clear, high-performing offerings

Tools:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit-learn: Random Forest & Linear Regression