E-commerce Return Analysis Project Report

# 1. Objective

The goal of this project is to analyze product return patterns in an e-commerce setting using Python, SQL, and Power BI. The focus is on identifying high-risk return categories, understanding geographical trends, and predicting the likelihood of returns to aid business decision-making.

# 2. Tools and Technologies

- Python (Data cleaning, model training)

- Pandas, NumPy (Data manipulation)

- Scikit-learn (Logistic Regression)

- SQL (MySQL) (Data storage and querying)

- Power BI (Visualization and Dashboarding)

# 3. Data Sources

a. Raw Data

- CSV file containing online retail transactional data.

b. Cleaned Data Files

- cleaned\_data.csv (after initial cleaning)

- cleanest\_data.csv (after final processing)

- high\_risk\_products.csv (model prediction output)

# 4. Data Preparation

Scripts used:

- clean\_data.py:

- Removed missing values and duplicates.

- Standardized column names.

- Saved cleaned data to cleaned\_data.csv.

- data import.py:

- Loaded cleaned data into MySQL using SQLAlchemy.

- Created derived features like return\_flag and category.

- Trained a logistic regression model to predict return probability.

- Exported high-risk products with return probabilities > 60%.

# 5. SQL Analysis

SQL script (data.sql) was used to:

- Create a database and table for cleaned data.

- Run aggregated queries by Description and Country:

- Total orders

- Total quantity sold

- Total revenue

# 6. Machine Learning Model

- Model Used: Logistic Regression

- Features: unitprice, quantity, category (One-Hot Encoded)

- Target: return\_flag (1 if quantity < 0, else 0)

- Evaluation:

- Train/Test split: 80/20 with stratification

- Model performance evaluated using a classification report (precision, recall, F1-score)

# 7. Key Metrics from Analysis

| Metric | Value |

|-------------------|-------------|

| Total Orders | ~8,000 |

| Total Returns | 402 |

| Return Rate | ~5% |

| High Risk Products| Exported if return risk > 60%

# 8. Power BI Visualizations

Based on main.pbix and report screenshots:

- Return Rate by Category:

- Categories like "ESSENTIAL", "PENS", and "DOGGY" had higher return rates.

- Total Returns by Country:

- Visualized on a global map using Power BI.

- Returns by Time (Year, Quarter, Month, Day):

- Seasonal trends observed in December months.

- Return Risk Dashboard:

- Heatmaps, bar graphs, and maps used to show spatial and categorical risk factors.

# 9. Conclusion

- A predictive model successfully estimated the likelihood of product returns.

- Certain categories and regions showed significantly higher return rates.

- Business teams can use the exported high\_risk\_products.csv to adjust policies, improve packaging, or enhance product descriptions.