# Return Analysis Project Report

## 1. Introduction

In the modern e-commerce landscape, efficient return management is a critical success factor for customer satisfaction and profitability. This project focuses on identifying high-risk products that are more likely to be returned and understanding the underlying factors contributing to product returns using customer behavior, product ratings, and service quality metrics.

The system also includes a prediction model and an interactive dashboard to support return analysis and decision-making.

---

## 2. Abstract

The objective of this project is to build an integrated return analysis framework using real-world-style datasets. The framework includes:

- **Data preprocessing** from 9 CSV sources (Customers, Orders, Products, Ratings, Returns, etc.)
- **High-risk product identification** using refund frequency and low customer ratings

- **Return prediction modeling** using machine learning
- **Dashboard visualization** to explore return patterns and KPIs

The system supports business teams in proactively minimizing returns and optimizing customer satisfaction.

---

## 3. Tools Used

- **Microsoft SQL Server Management Studio** – Data querying and joining from multiple sources
- **Python (Pandas, Sklearn, Matplotlib)** – Data preprocessing, model building
- **Jupyter Notebook / Anaconda** – Script development
- **Power BI / Excel Dashboard** – Visualization layer
- **CSV/Excel Files** – Source datasets

---

## 4. Steps Involved in Building the Project

1. **Data Collection:**
   Nine CSV files containing customer, order, product,

transaction, delivery, and subscription data were imported.

2. **Data Integration:**
   SQL joins were used to combine relevant tables like Orders, Products, Ratings, Returns_Refund, and Transactions.
3. **Data Cleaning & Preprocessing:**
   Python scripts were written to handle missing values, normalize rating scales, and categorize refund reasons.
4. **High-Risk Product Identification:**
   Products were marked "high risk" if they had:
   - Frequent refunds (via Returns_Refund)
   - Low product and delivery ratings (via Ratings.csv)
5. **Return Prediction Model:**
   A machine learning model (e.g., Logistic Regression or Random Forest) was developed to predict whether an order is likely to be returned based on features like product rating, delivery rating, and order quantity.
6. **Dashboard Creation:**
   An interactive dashboard was created to display:
   - Top returned products

- Rating distributions
- Refund frequency over time
- Key performance indicators

---

## 5. Conclusion

This project successfully integrates data engineering, machine learning, and visualization to address the challenge of product returns in e-commerce. By identifying high-risk products and enabling predictive insights, the project provides value to inventory management, customer service, and product development teams.

The tools and pipeline created can be extended to live data systems and enhanced with deeper NLP-based sentiment analysis from customer feedback in future iterations.