Capstone Project: E-commerce Sales Analysis

Project Overview

This project performs an **end-to-end data analysis of an e-commerce platform** (simulating Amazon sales data). The analysis covers the entire workflow: from raw messy data cleaning, exploratory data analysis, advanced analysis, and insights generation to actionable business recommendations.

Goals and Objectives:

- Analyze sales data to identify trends, top products, and revenue distribution.
- Detect and handle inconsistencies, missing values, and formatting issues in raw data.
- Understand customer behavior and preferences.
- Generate actionable recommendations to improve sales and marketing strategies.
- Create visualizations and reports suitable for business stakeholders.

Setup Instructions

Environment Requirements:

- Python 3.9+
- Libraries: pandas, numpy, matplotlib, seaborn, openpyxl

Installation Guide:

1. Install Python from python.org.
2. Clone or download the project repository.
3. Navigate to the project folder and install dependencies:

pip install -r requirements.txt

4. Ensure the data is placed in the data/ folder:
    - data/raw_data.xlsx (raw dataset)
5. Run the notebooks/scripts in order:
    - 1_data_cleaning.py → cleans raw data
    - 2_eda.py → performs exploratory data analysis
    - 3_analysis.py → advanced analysis and actionable insights

---

Code Structure

```
capstone_project/
|
 ├– data/
|    ├–amazon_data.xlsx
|    └– cleaned_data.xlsx
|
 ├– notebooks/
|    ├– data_cleaning.py    # Data cleaning & preprocessing
|    ├– 2_eda.py            # Exploratory Data Analysis
|    └– 3_analysis.py       # Advanced analysis & recommendations
|
 ├– reports/
|    ├– executive_summary.pdf
|    └– technical_report.pdf
|
 ├– presentations/
|    └– business_presentation.pptx
|
 ├– README.md
└– requirements.txt
|
 ├– Visual/
|    ├–report
|    └– EDA
|
```

Visual Documentation

The project includes **graphs and plots** generated by Python scripts to demonstrate insights:

1. **Distribution of numerical features:** histograms for price, quantity, discount, tax, and shippingcost.
2. **Categorical distributions:** countplots for category, subcategory, region, paymentmethod, and orderstatus.
3. **Correlation heatmap:** showing relationships between numerical features.
4. **Revenue analysis:** total revenue by category, region, and top products.
5. **Monthly sales trend:** line plot showing sales over time.
6. **Category vs Region heatmap:** highlights top-performing regions and categories.
7. **Discount vs Sales scatterplot:** visualizes the effect of discounts on revenue.
8. **Top customers by revenue:** identifies high-value customers for targeted marketing.

Technical Details

Algorithms & Techniques Used:

- **Data Cleaning:**
    - Handle missing values with default values or median imputation.
    - Convert data types (Price, Quantity, Discount, Tax) to numeric.
    - Standardize text fields (Region, Category, PaymentMethod).
    - Remove duplicates and correct negative prices.
- **Exploratory Data Analysis (EDA):**
    - Histograms, countplots, scatterplots, and heatmaps.
    - Aggregations by category, region, product, and customer.
- **Advanced Analysis:**
    - Top products by quantity sold and revenue.
    - Monthly revenue trends and seasonality.
    - Revenue correlation with discounts and shipping costs.
    - Identification of upsell opportunities based on high quantity but low revenue.

Data Structures:

- **DataFrames (pandas)** used throughout for structured data manipulation.
- Pivot tables for multi-dimensional revenue aggregation.

Architecture:

- Modular approach: each Python script/notebook performs a specific workflow stage.
- Cleaned data stored as cleaned_data.xlsx for downstream analysis.

Testing Evidence

Example Test Cases and Validations:

1. **Missing values:** Verified that null CustomerID and CustomerName are replaced with 'UNKNOWN'.
2. **Data type conversions:** Checked that all numeric fields (Price, Quantity, Discount, Tax) are numeric.
3. **Negative prices:** All negative values replaced with median price.
4. **Duplicate removal:** Verified no duplicate OrderID remains.
5. **Revenue calculation:** total_sales = price * quantity * (1-discount) + tax + shippingcost.
6. **Consistency checks:** Regions, categories, and payment methods standardized.
7. **Insights validation:** Top products, regions, and customers checked against aggregated sums.

Summary of Deliverables

- **Executive Summary:** high-level business insights (1 page).
- **Technical Report:** detailed methodology, analysis, and visualizations (5–10 pages).
- **Presentation:** 10–15 slides for business stakeholders.
- **Code Repository:** clean, modular, and documented Python scripts.

Outcome: The project demonstrates **data-driven business decision-making**, highlights key insights, and proposes actionable recommendations for an e-commerce platform.