Online Marketplace Sales Analysis Project

Course: Data Analysis & Python Programming
Points: 60
Duration: 10 days
Individual Project

Project Overview

You have been hired as a **Data Analyst** for **GlobalMarket Analytics**, a consulting firm that specializes in e-commerce performance optimization. Your client is a major online marketplace platform that operates globally across multiple product categories and payment systems.

The marketplace management team needs comprehensive insights into their sales performance, regional trends, product category effectiveness, and customer payment preferences to make strategic decisions for the upcoming fiscal year.

Your mission is to analyze their transaction data and provide actionable recommendations that will help optimize their platform performance, identify growth opportunities, and improve operational efficiency.

Dataset Information

Dataset Source:

- **Download Link:** https://www.kaggle.com/datasets/shreyanshverma27/online-sales-dataset-popular-marketplace-data
- **Dataset Name:** Online Sales Dataset - Popular Marketplace Data
- **Size:** Approximately 100,000+ transaction records
- **Format:** CSV file
- **License:** Public Domain

Dataset Description:

This dataset contains global online sales transactions across various product categories from a popular marketplace platform. It represents realistic e-commerce sales data with natural variations and patterns that mirror real-world business scenarios.

Column Descriptions:

Column Name	Data Type	Description	Example Values
Order_ID	String	Unique identifier for each transaction	"ORD-2023-001", "ORD-2023-002"
Date	Date	Transaction date (YYYY-MM-DD format)	"2023-01-15", "2023-03-22"
Category	String	Main product category	"Electronics", "Clothing", "Books", "Home & Garden"
Product_Name	String	Specific product sold	"Wireless Bluetooth Headphones", "Cotton T-Shirt"
Quantity	Integer	Number of units sold in transaction	1, 2, 5, 10
Unit_Price	Float	Price per individual unit (USD)	29.99, 15.50, 199.00
Total_Price	Float	Total revenue from transactions (Quantity × Unit_Price)	59.98, 31.00, 995.00
Region	String	Geographic region of sale	"North America", "Europe", "Asia", "South America"
Payment_Method	String	Customer payment method	"Credit Card", "PayPal", "Bank Transfer", "Digital Wallet"

Data Characteristics:

- **Time Period:** Covers 12+ months of transaction data
- **Geographic Coverage:** Global sales across multiple regions
- **Product Diversity:** 10+ main categories with hundreds of unique products
- **Natural Data Issues:** Contains realistic data quality challenges including:
  - Missing values in some fields
  - Inconsistent date formats
  - Potential duplicate records
  - Outliers in some fields

    - Seasonal sales patterns

- **Business Complexity:** Includes various payment methods, international transactions, and different product categories

---

Project Requirements

1. Initial Data Investigation (5 points)

Conduct a thorough initial examination of the raw dataset to understand its structure, content, and basic characteristics before proceeding with any analysis.

- **Dataset Loading & Structure**: Load the dataset and examine its dimensions, column names, and basic structure
- **Data Types Analysis**: Identify the data types of each column and assess if they align with expected formats
- **Initial Content Examination**: Review sample records to understand the actual data content and formats
- **Basic Statistical Overview**: Generate preliminary descriptive statistics for numerical columns
- **Categorical Variable Exploration**: Examine unique values and frequency distributions in categorical columns
- **Data Volume Assessment**: Understand the scale and scope of the dataset (number of transactions, time period covered, etc.)
- Document first impressions and initial observations about the data

Deliverable: Initial data investigation summary with key observations about dataset characteristics, structure, and preliminary insights.

Note: Type your observations in a separate markdown in your notebook.

---

2. Data Quality Assessment & Cleaning (10 points)

Conduct a comprehensive evaluation of the dataset's integrity and implement appropriate data cleaning procedures.

Focus Areas:

- **Missing Value Analysis**: Identify the number of missing data across Order_ID, Region, Payment_Method, and other fields, make a dataframe with these columns [column name, number of missing, the percentage of these missing out of the whole column]
- **Make some visualizations**: visualize the missing data using missing_no library for the fields.
- **Determine the missing data mechanism**: for the data you identified the missing in them explain what's the missing data mechanism **(MCAR, MNAR, MAR),** and why you see this is the best mechanism, decide what's the best way to handle the missings in the columns.
- **Clean missing data:** Code your decisions about the missing data.
- **Data Type Validation**: Ensure proper data types for dates, prices, quantities, and categorical variables
- **Duplicate Detection**: Identify and handle potential duplicate Order_IDs or suspicious transaction patterns
- **Outlier Investigation**: Analyze extreme values in the fields and determine what these fields are.
- **Outlier Handling**: After determining what the columns that have outliers are, choose the best way to handle these outliers, and explain why you chose this way.
- **Data Consistency**: Verify that Total_Price calculations align with Unit_Price × Quantity
- **Date Format Standardization**: Ensure consistent date formatting and handle any parsing issues

Deliverable: Comprehensive data quality report documenting all issues found and cleaning decisions made.

---

3. Exploratory Data Analysis (10 points)

Perform thorough exploratory analysis to understand marketplace dynamics and transaction patterns using the original dataset columns only. Do not create new features in this section - analyze the data as-is.

Required Analysis Areas:

Temporal Analysis (using Date column):

- **Daily/Weekly/Monthly sales patterns:** Plot total revenue and transaction count over time to identify which days of the week, weeks of the month, or months have higher sales activity
- **Seasonal variations and peak periods:** Analyze the Date column to determine which months or seasons generate the highest revenue and transaction volume
- **Sales trend identification:** Create line charts showing revenue progression over time to identify if sales are increasing, decreasing, or stable throughout the dataset period

Product Category Performance (using Category, Unit_Price, Quantity, Total_Price):

- **Revenue distribution across categories:** Calculate and visualize what percentage of total marketplace revenue each product category generates (Electronics vs. Clothing vs. Books, etc.)
- **Average order values by category:** Compare the mean Total_Price for transactions in each category to identify which categories generate higher-value purchases
- **Quantity patterns by category:** Analyze typical Quantity values purchased for each category - do customers buy Electronics in single units but Clothing in bulk?
- **Category popularity rankings:** Rank categories by total number of transactions, total revenue, and average transaction value to identify top performers

Geographic Analysis (using Region, Total_Price, Category):

- **Regional sales performance comparison:** Compare total revenue, average transaction value, and transaction count across North America, Europe, Asia, and South America
- **Revenue distribution across regions:** Calculate what percentage of total marketplace revenue comes from each geographic region
- **Regional category preferences:** Analyze which product categories are most popular in each region - does Asia prefer Electronics while Europe prefers Clothing?

Payment Method Analysis (using Payment_Method, Total_Price, Region):

- **Payment method adoption rates:** Calculate what percentage of transactions use Credit Card vs. PayPal vs. Bank Transfer vs. Digital Wallet
- **Average transaction values by payment type:** Compare mean Total_Price across different payment methods to identify if certain payment types correlate with higher spending
- **Regional payment preferences:** Analyze which payment methods are preferred in each geographic region

Price and Quantity Insights (using Unit_Price, Quantity, Total_Price, Category):

- **Price distribution analysis across categories:** Create histograms and box plots showing Unit_Price ranges for each category to understand pricing structures
- **Quantity purchase patterns:** Analyze the distribution of Quantity values to identify if customers typically buy single items, small quantities, or bulk amounts
- **Revenue concentration analysis:** Identify which percentage of transactions account for the majority of revenue (e.g., do 20% of transactions generate 80% of revenue?)

Cross-Variable Relationships:

- **Price vs. Quantity correlation:** Investigate if higher Unit_Price products are purchased in smaller or larger quantities
- **Regional spending patterns:** Compare average Total_Price across different regions
- **Category-Payment method relationships:** Analyze if certain categories are more likely to be paid for with specific payment methods

Requirements:

- Use only original dataset columns - Order_ID, Date, Category, Product_Name, Quantity, Unit_Price, Total_Price, Region, Payment_Method
- Create meaningful visualizations for each analysis area (bar charts, line plots, pie charts, histograms, box plots)
- Calculate relevant statistical measures (mean, median, standard deviation, percentiles, counts, percentages)
- Identify and investigate interesting patterns or anomalies in the existing data

- Document findings with clear interpretations of what each analysis reveals about the marketplace

Note: This section focuses on understanding and exploring the existing data structure and patterns. You will create new features and derived variables in Section 4 (Feature Engineering)

---

4. Feature Engineering & Data Transformation (8 points)

Create new variables and transform existing data to enhance analytical capabilities.

Required Feature Categories:

Required Analysis Areas:

Temporal Features:

- **Extract day of week, month, quarter from Date column**: Break down the transaction dates into separate components to analyze patterns - for example, create new columns showing whether transactions occurred on Monday vs. Friday, in January vs. December, or in Q1 vs. Q4. This helps identify if certain days, months, or quarters have consistently higher or lower sales.
- **Create seasonal indicators (Spring, Summer, Fall, Winter)**: Group months into seasons to analyze seasonal purchasing behavior - assign March-May as Spring, June-August as Summer, September-November as Fall, and December-February as Winter. This reveals whether certain product categories sell better during specific seasons.
- **Identify weekdays vs. weekends**: Create a binary indicator showing whether each transaction occurred on a weekday (Monday-Friday) or weekend (Saturday-Sunday) to understand shopping behavior patterns and potentially optimize marketing strategies for different parts of the week.
- **Calculate days since first transaction (customer lifecycle indicators)**: Determine how many days have passed since the very first transaction in your dataset to each subsequent transaction, creating a timeline that shows marketplace maturity and can help identify early vs. late adoption patterns in different regions or categories.

Business Metrics:

- **Price per unit categories (Low, Medium, High based on quartiles)**: Divide all Unit_Price values into three groups using statistical quartiles - products in the bottom 25% of prices are "Low," middle 50% are "Medium," and top 25% are "High." This categorization helps analyze purchasing behavior across different price ranges.
- **Order size categories (Single item, Small bulk, Large bulk)**: Group transactions by Quantity values to understand purchase patterns - for example, Quantity = 1 could be "Single item," 2-5 could be "Small bulk," and 6+ could be "Large bulk." This reveals whether customers typically buy individual items or prefer bulk purchases.
- **Revenue tiers for transactions**: Categorize Total_Price values into meaningful business segments such as "Budget" (lowest 33% of transaction values), "Standard" (middle 33%), and "Premium" (highest 33%) to identify spending patterns and high-value customer behavior.
- **Product category groupings (if applicable)**: If there are many individual product categories, consider grouping related categories together for analysis - for example, group "Electronics," "Computers," and "Mobile Phones" into "Technology Products," or combine "Clothing," "Shoes," and "Accessories" into "Fashion Items" to simplify analysis and identify broader market trends.

Geographic Features:

- **Revenue concentration by region**: Calculate what percentage of total marketplace revenue comes from each geographic region to identify dominant markets
- **Regional market penetration metrics**: Develop measures showing how well the marketplace performs in each region (e.g., average revenue per transaction by region, transaction volume by region)

Advanced Features:

- **Average selling price by product category**: Calculate the typical price point for each product category to understand pricing tiers across the marketplace
- **Category performance relative to overall marketplace**: Create metrics showing how each product category performs compared to the overall marketplace average (e.g., above/below average revenue, above/below average transaction volume)

- **Payment method efficiency metrics**: Develop measures of payment method performance such as average transaction size by payment type, or adoption rates across different regions
- **Transaction complexity scores**: Create indicators of transaction characteristics such as high-value vs. low-value purchases, single-item vs. multi-item orders, or premium vs. budget product transactions.

Feature Transformation:

- **Categorical Data Conversion**: Transform text-based columns (Category, Region, Payment_Method) into categorical data types to optimize memory usage and enable proper statistical analysis
- **Drop product name column**
- **Numerical Encoding**: Convert categorical variables into numerical representations suitable for mathematical operations and statistical modeling (e.g., Region categories become numerical codes, Payment_Method types become numerical identifiers)
- **Data Type Optimization**: Ensure all columns have appropriate data types for analysis, converting string objects to their proper categorical or numerical formats as needed

> **Additional requirement**: Make some visualizations about the new features that you created to support analysis and make the stakeholders understand your findings.

---

5. Statistical Analysis & Hypothesis Testing (7 points)

Formulate and test relevant business hypotheses using appropriate statistical methods.

Required Hypothesis Tests:

Regional Performance Comparison:

- Test if mean transaction values differ significantly across regions
- Investigate if product category preferences vary by region

Payment Method Analysis:

- Test if average order values differ by payment method

- Analyze if payment method adoption is uniform across regions

Product Category Insights:

- Test for significant differences in average prices across categories
- Investigate seasonal effects on category performance

Requirements:

- Clearly state null and alternative hypotheses
- Choose and justify appropriate statistical tests (t-tests, ANOVA, chi-square, etc.) (search online for the proper test based on the above use cases)
- Report the output of test.

---

6. Advanced Analytics & Business Insights (10 points)

Apply analytical techniques to generate actionable business intelligence and strategic recommendations.

Choose 3 of the following approaches:

Customer Segmentation (Transaction-Based):

- **Segment transactions by order value, quantity, and frequency patterns**: Group transactions into meaningful categories such as "Budget Shoppers" (low order values), "Bulk Buyers" (high quantities), or "Premium Customers" (high-value purchases). Analyze what percentage of transactions fall into each segment.
- **Analyze payment method preferences by segment**: Investigate whether high-value customers prefer credit cards while budget shoppers use digital wallets, or if regional preferences affect payment choices within each segment.
- **Identify high-value transaction characteristics**: Determine what makes a transaction valuable - is it specific product categories, certain regions, particular times of year, or specific payment methods? Create a profile of your most profitable transactions.
- **Create customer personas based on purchasing behavior**: Develop detailed profiles like "Electronics Enthusiast" (frequently buys electronics with high order values) or "Regional Bulk Buyer" (purchases large quantities from specific geographic areas).

Revenue Optimization Analysis:

- **Identify underperforming product categories by region**: Find which product categories generate below-average revenue in specific regions - for example, if "Books" performs poorly in Asia but well in Europe, investigate why and propose solutions.
- **Analyze pricing patterns and opportunities across categories**: Compare price ranges within categories and across regions to identify pricing inconsistencies or opportunities - are Electronics priced similarly across all regions, or are there market-specific pricing advantages?
- **Compare average order values across different dimensions**: Calculate and compare average transaction values by category, region, payment method, and time periods to identify the most and least profitable combinations.
- **Recommend pricing strategies for different markets**: Based on your analysis, suggest specific pricing adjustments - should certain categories be priced higher in regions where they're popular, or should underperforming categories be discounted to increase volume?

Comparative Performance Analysis:

- **Benchmark category performance against marketplace averages**: Calculate overall marketplace metrics (average transaction value, average quantity per order) and compare how each product category performs relative to these benchmarks - which categories exceed expectations and which fall short?
- **Identify top and bottom performing products within categories**: Within each category, rank individual products by revenue, quantity sold, or transaction frequency to identify star performers and products that may need attention or discontinuation.
- **Analyze regional competitive advantages**: Determine which regions excel at selling specific categories - does North America dominate Electronics sales while Europe leads in Clothing? Identify each region's strengths and weaknesses.
- **Compare payment method efficiency and adoption rates**: Evaluate which payment methods generate higher average transaction values, are adopted most frequently, and vary by region - are credit card users spending more than PayPal users?

Business Intelligence Metrics Development:

- **Design key performance indicators (KPIs) for the marketplace**: Create measurable metrics such as "Revenue per Region per Month," "Category Market Share," "Average

Customer Lifetime Value," or "Payment Method Conversion Rates" that executives could monitor regularly.

- **Create comprehensive business metrics and benchmarks**: Establish baseline performance standards like "Target: 15% quarterly revenue growth" or "Benchmark: Electronics should represent 30% of total sales" against which future performance can be measured.
- **Develop actionable insights summary with supporting evidence**: Create executive-level summaries such as "Revenue Opportunity: Increasing Electronics marketing in South America could boost revenue by 12% based on regional underperformance analysis."
- **Propose monitoring strategies for ongoing business health**: Recommend which metrics should be tracked daily, weekly, or monthly, and suggest alert thresholds that would indicate when business performance requires immediate attention.

Cross-Dimensional Analysis:

- **Investigate relationships between multiple variables simultaneously**: Explore complex questions like "Do customers in Europe who use Credit Cards for Electronics purchases spend more during winter months?" or "Are bulk Electronics purchases more common in Asia during specific seasons?"
- **Analyze interaction effects between region, category, and payment methods**: Determine if certain combinations perform exceptionally well or poorly - for example, do Credit Card purchases for Home & Garden products in North America generate higher profits than other combinations?
- **Identify unexpected patterns and business opportunities**: Look for surprising insights such as "Digital Wallet users in South America prefer high-value Electronics purchases" or "Books category peaks during different seasons in different regions," which could inform marketing strategies.
- **Create multi-factor performance matrices**: Develop comprehensive comparison tables showing how different combinations of region, category, and payment method perform across various metrics like average order value, transaction frequency, and total revenue contribution.

Technical Requirements

Core Tools & Libraries that you can use (not mandatory – it’s up to you):

- **Data Manipulation:** pandas, numpy
- **Visualization:** matplotlib, seaborn, plotly (choose appropriate tools)
- **Statistics:** scipy.stats, statsmodels
- **Additional:** Any relevant libraries for advanced analysis (scikit-learn, etc.)

Code Quality Standards:

- Well-documented code with clear comments explaining methodology
- Typing your functions
- Efficient pandas operations and numpy computations
- Proper error handling data quality issues
- Clear variable naming and code organization

Visualization Requirements:

- Professional-quality charts with proper labels, titles, and legends
- Appropriate chart types for different data types
- Color schemes that enhance readability
- Interactive visualizations where beneficial
- Clear annotations and explanations

Deliverables

1. Jupyter Notebook Analysis (50 points)

Complete analytical notebook containing:

- Initial data investigation and structure analysis
- Data loading and quality assessment
- All code, outputs, and visualizations
- Clear section organization with markdown documentation
- Statistical analysis with proper interpretation

- Advanced analytics implementation
- Business insights and pattern identification

2. Executive Business Report (10 points)

3–5-page professional report including:

- **Executive Summary:** Key findings and recommendations
- **Market Performance Overview:** Regional and category insights
- **Strategic Recommendations:** Actionable business suggestions
- **Data Limitations:** Quality issues and analysis constraints
- **Next Steps:** Proposed follow-up analyses and data collection needs