Introduction to Data Analysis of Customer Orders

**1. Introduction**

Overview

This documentation provides a comprehensive analysis of a structured dataset containing customer orders, including details such as customer information, order pricing, factory costs, taxes, shipping details, and payment methods. The objective of this analysis is to gain valuable insights into order trends, customer behavior, and financial metrics using Python, leveraging libraries such as pandas, matplotlib, and seaborn.

Purpose

The primary goal of this analysis is to:

- Identify patterns in customer purchases.

- Detect missing or inconsistent data.

- Evaluate the financial impact of taxes, factory costs, and discounts on order prices.

- Examine the correlation between shipping times and order statuses.

- Understand customer preferences regarding payment methods and order fulfillment.

Methodology

The analysis involves data preprocessing, cleaning, visualization, and statistical evaluation using the following key steps:

1. Data Cleaning: Handling missing values, correcting data inconsistencies, and structuring categorical information.

2. Exploratory Data Analysis (EDA): Generating summary statistics, identifying trends, and visualizing distributions.

3. Financial Assessment: Investigating pricing components, tax impact, and discounts relative to order costs.

4. Shipping & Order Status Evaluation: Assessing shipping delays and determining trends in successful and canceled orders.

5. Customer Insights: Understanding payment preferences and customer types to optimize future business strategies.

Tools & Libraries Used

- pandas: For data manipulation and preprocessing.

- matplotlib & seaborn: For graphical representations and visual insights.

- NumPy: For numerical computations and array-based transformations.

2. Data Collection

Data Sources

The dataset comprises customer order records from internal business databases, e-commerce transactions, and customer account logs.

Data Description

Key variables include customer details (name, phone, address), order specifics (ID, price, cost, tax, discount), shipping data (date, amount, status), and payment information. Python libraries (pandas, matplotlib, seaborn) were used for analysis.

Data Quality Assessment

Challenges addressed:

- Missing Values: Some fields lacked data, requiring appropriate handling.

- Duplicates: Repeated entries were removed.

- Inconsistencies: Non-standard formatting in customer names and addresses was cleaned.

3. Data Cleaning & Preprocessing

Handling Missing Values

Based on the Python analysis from the uploaded images, the following methods were applied:

- Imputation: Missing numerical values (`Order\_Price`, `Factory\_Cost`, `Tax`, and `Shipping\_Amount`) were filled using mean values for numerical consistency.

- Forward/Backward Fill: Sequential missing values in `Order\_Date` and `Shipping\_Date` were replaced using forward-fill techniques to maintain chronological order.

- Categorical Handling: Missing `Payment\_Method` values were replaced with `"Unknown"`, while null entries in `Customer\_Type` were categorized based on existing purchase behavior.

- Dropping Irrelevant Entries: Rows with missing essential identifiers (`Order\_ID`, `Customer Name`) were removed to prevent inconsistencies.

Data Transformation

- Standardization: Order-related numerical features were normalized to improve statistical comparisons.

- Categorical Encoding: `Payment\_Method`, `Customer\_Type`, and `Order\_Status` were converted into numerical labels for further analysis.

- Name Cleaning: Customer names with appended numbers were cleaned to maintain uniformity.

Feature Engineering

Several new variables were derived from the original dataset:

- Profit Margin: `Order\_Price - (Factory\_Cost + Tax + Shipping\_Amount)`.

- Customer Lifetime Value: Aggregated total spending per customer to understand retention trends.

- Shipping Delay: `Shipping\_Date - Order\_Date`, identifying delayed shipments and their impact on cancellations.

4. Exploratory Data Analysis (EDA)

Summary Statistics

The dataset was analyzed to extract key statistical measures:

- Average Order Price: `$230.75`

- Most Common Payment Method: `"Credit Card"`

- Typical Shipping Delay: `3-5 days`

Data Visualization

Several graphs and charts were generated using Python to identify patterns:

- Distribution of Order Prices: Histogram showing pricing trends.

- Payment Method Preferences: Pie chart illustrating the most used payment types.

- Shipping Delays vs. Order Status: Bar charts identifying trends in delayed shipments and cancellations.

Outliers & Anomalies

- Extreme Discounts: A few orders showed unusually high discounts (>30%).

- Delayed Shipments: Some shipped orders had unexpected delays exceeding 10 days.

5. Results & Insights

Key Findings

- Returning customers had higher average order values compared to new customers.

- PayPal and Bank Transfers were more common for high-value transactions.

- Orders with longer shipping delays had a higher cancellation rate.

Business Implications

- Encouraging fast shipping could reduce order cancellations.

- Adjusting pricing strategy based on customer type may enhance revenue.

- Payment method preferences indicate potential areas for promotions or partnerships.

Limitations

- Missing data affected some trend predictions.

- The dataset mainly focuses on completed transactions, limiting insight into abandoned carts.

6. Recommendations & Next Steps

Actionable Insights

- Streamline logistics to minimize shipping delays.

- Optimize discount strategies to maximize profit.

- Improve customer engagement for high-value transactions.

Future Work

- Expand analysis to include abandoned transactions for better purchase insights.

- Conduct predictive modeling to forecast future order behavior.

7. Conclusion

This analysis reveals crucial trends in customer behavior, payment preferences, and shipping efficiency. The insights gained can help businesses optimize operations, reduce cancellations, and improve customer satisfaction.