1. Data Cleaning and Transformation Report

a) Data Collection and Merging:

- **Datasets:** Integrated data from three separate e-commerce datasets.
- **Final Dataset**: Merged the individual datasets into a consolidated file named 'DATAAA.csv'.

b) Data Import and Initial Exploration:

- **Dataset:** Imported using `pd.read_csv` from 'DATAAA.csv'.
- **Initial Data Check:**
  Displayed the first few rows and columns of the dataset.
  Checked for duplicates and missing values.

c) Duplicate Handling:

- Identified and reviewed duplicates in `product_id` and `seller_id`.
- Removed duplicates based on `product_id` and `seller_id`, keeping the first occurrence.
- Ensured uniqueness of `product_id` and `seller_id` after removal.

d) Missing Values Handling:

- **Numeric Columns:** Filled missing values with the mean for numeric columns such as `price`, `freight_value`, and others.
- **Categorical Columns:** Filled missing values with 'unknown' or the mode for categorical columns like `seller_city`, `order_item_id`, and others.
- **Additional Specific Columns:** appropriate imputation methods for columns like `market_share_product`, `zip_code_prefix`, `payment_sequential`, etc.

e) Outlier Detection and Treatment:

- **IQR Method:**
  Calculated the Interquartile Range (IQR) for numeric columns and removed outliers by setting them to NaN
- **Z-Score Method:**
  Calculated Z-scores for numeric columns and identified outliers with Z-scores greater than 3.

f) Normalization and Standardization

- **Normalization:**
  Applied Min-Max scaling to numerical columns to normalize the values within a range [0, 1].
- **Standardization:**
  Standardized numerical columns using `StandardScaler` to have a mean of 0 and a standard deviation of 1.

g) Categorical Data Encoding

- **One-Hot Encoding:**
  Applied one-hot encoding to categorical columns including `order_id`, `customer_state`, `product_id`, `seller_id`, and others.

h) Additional Metrics Calculation

- **Total Sales per Transaction:**
  Calculated as the sum of `price` and `freight_value` for each `TransactionID`.
- **Average Transaction Value:**
  Computed as the total sales per transaction divided by the number of items per transaction.
- **Normalization of Metrics:**
  Standardized and then normalized the `total_sales_per_transaction` and `average_transaction_value` columns.

i) Final Data Preparation

- **Columns Drop:**
  Dropped unnecessary columns such as `review_comment_title` to streamline the dataset.
- **Final Data Check:**
  Displayed the updated DataFrame with the latest transformations applied.

2. Dimensionality Reduction with PCA:

a) Application of PCA:

- **Methodology:**
  - **Data Standardization**: The numerical features were standardized using `StandardScaler` to ensure that each feature contributes equally to the PCA. This step is crucial as PCA is sensitive to the scale of the data.
  - **PCA Initialization and Fitting:** PCA was applied to the standardized data. Initially, all principal components were extracted to determine how many components to retain based on the explained variance.
  - **Component Selection**: After fitting PCA, the explained variance ratio was analyzed to decide on the number of principal components. For simplicity and interpretability, 2 components were chosen.
- **Reason for Component Choice:**
  - The choice of 2 components was driven by the need for visualization and interpretability. These two components often capture the most significant variance in the data, allowing for effective visualization and insight extraction.

b) Visualization of Principal Components

- **Scatter Plot:**
  - **Visualization Technique**: A scatter plot of the first two principal components (PC1 and PC2) was created.
  - **Insights**:
  - **Clustering:** The scatter plot may reveal natural clusters or groupings in the data. For example, if certain customer segments or product types cluster together, this could indicate similarities in purchasing patterns or behaviors.
  - **Separation:** The plot also helps in identifying any separation between categories, such as different customer states or product categories.

- **Visualization of PCA with Customer Segments:**
  - **Hue-based Visualization**: To understand how customer segments are distributed in the PCA space, a scatter plot was overlaid with customer state as the hue.
  - **Insights:**
  - **Segment Distribution**: This visualization can highlight how different customer segments are distributed along the principal components. It can reveal whether certain segments are more clustered or dispersed.
  - **Category Separation:** It helps in observing if there is a clear separation between different customer states or other categorical variables.

c) Impact of Dimensionality Reduction

- **Discussion:**
  - **Loss of Information:** PCA reduces dimensionality, which can lead to a loss of some information, particularly in higher dimensions. However, the goal is to retain the most critical variance with fewer components.
  - **Understanding Data Structure**: By reducing dimensions, PCA helps in simplifying the data structure, making it easier to visualize and interpret. It provides a more manageable view of complex datasets and can reveal underlying patterns and relationships that might be obscured in high-dimensional space.

3. Customer Lifetime Value (CLV)

a) Calculation of CLV

- **Methodology:**
  - **Average Purchase Value:** Calculated as the mean price per customer.
  - **Purchase Frequency:** Determined by counting the number of orders per customer.
  - **Retention Rate**: A fixed retention rate (e.g., 0.8) was assumed for the calculation.
  - **CLV Formula:** CLV was computed using the formula: \(\text{CLV} = \text{Average Purchase Value} \times \text{Purchase Frequency} \times \text{Retention Rate} \).

- **Insights:**
  - CLV provides a measure of the total value a customer is expected to bring to a business over their lifetime. This metric is crucial for understanding customer profitability and for strategic decision-making in marketing and customer retention.

b) Visualization of CLV Distribution

- **Bar Plot of Average CLV by Segment:**
  - **Visualization Technique:** A bar plot was created to show the average CLV across different customer segments.
  - **Insights:**
  - **Segment Performance:** This visualization helps identify which customer segments are most valuable and which have the highest average CLV.
  - **Strategic Focus:** Segments with higher CLV can be targeted for more personalized marketing efforts.
- **Box Plot of CLV Distribution:**
  - **Visualization Technique**: A box plot was used to visualize the distribution of CLV across different customer segments.
  - **Insights:**
  - **Distribution Spread:** This plot shows the variability and distribution of CLV within each segment, highlighting any outliers and the range of values.
  - **Comparative Analysis**: It allows for a comparison of CLV distributions across segments, providing insight into which segments have more consistent or variable customer values.

4. What-if Analysis

a) Impact of Changes in Transaction Value

- **Methodology:**
  - **Scenario Analysis:** A what-if analysis was conducted by varying the average transaction value and observing its impact on CLV.
  - **Visualization:** The results were visualized using line plots to show how changes in the transaction value affect the average CLV.
  - **Insights:**
  - **Sensitivity to Changes:** The analysis demonstrates how sensitive CLV is to changes in average transaction value. A significant increase in transaction value generally leads to a higher CLV.
  - **Strategic Implications:** This information is useful for understanding how changes in pricing or sales strategies might impact overall customer value and profitability.

5. Insights and Narrative:

a) Key Insights on Global E-Commerce Trends and Their Impact on Traditional Retail:

- **Rapid E-commerce Growth:**
  - E-commerce is growing much faster than traditional retail, showing a big shift in consumer shopping behavior towards online platforms.
- **Consumer Preference for Convenience:**
  - Shoppers increasingly prefer the convenience of online shopping, especially in regions with strong internet access.
- **Decline in Traditional Retail:**
  - Many brick-and-mortar stores, particularly in sectors like fashion and electronics, are seeing a drop in sales as more customers shop online.
- **Omnichannel Success:**
  - Retailers who combine online and offline shopping experiences (e.g., offering in-store pickup for online orders) are performing better than those sticking to physical stores only.
- **Importance of Fast Delivery:**
  - Quick and reliable delivery options have become key for e-commerce success, as traditional stores struggle to meet these expectations.

b) Narrative:

- **The Changing Face of Retail**

  The retail world is being reshaped by the rapid rise of e-commerce. More and more consumers are choosing the convenience of online shopping, leaving many traditional retailers to struggle. However, those who have embraced both digital and physical channels are finding new ways to thrive. Fast shipping, easy returns, and innovative services like same-day delivery are key to the growth of e-commerce platforms, making it difficult for traditional stores to compete.

  In this new era, retail is no longer limited to physical spaces but exists across multiple digital touchpoints. Retailers and platforms that adapt to this change will continue to succeed, while those who resist may struggle to stay relevant.

c) Actionable Recommendations

- **For Traditional Retailers:**
  - Go Omnichannel: Integrate online and offline shopping experiences to stay competitive.
  - Local Advantage: Use your physical presence to offer fast delivery or unique in-store experiences that online stores can’t provide.

    - Adopt In-Store Technology: Use tools like virtual shopping assistants or augmented reality to enhance the in-store experience.

- **For E-commerce Platforms:**
    - Improve Delivery: Keep investing in logistics for faster delivery times to keep customers happy.
    - Focus on New Markets: Expand into emerging markets where e-commerce is still growing.
    - Sustainability: Incorporate eco-friendly packaging and delivery options to meet the growing demand for sustainability.

- **For Policymakers:**
    - Improve Digital Infrastructure: Invest in better internet access, especially in underserved areas.
    - Reskill Workers: Create programs to train workers moving from traditional retail to digital and e-commerce roles.
    - Ensure Fair Competition: Regulate to make sure traditional retailers aren’t unfairly disadvantaged compared to e-commerce giants.