**Customer Data:**

| **customer\_id** | **age** | **gender** | **annual\_income** | **spending\_score** | **region** | **num\_purchases** |
| --- | --- | --- | --- | --- | --- | --- |
| CUST1000 | 56 | Male | 50939 | 7 | North | 39 |
| CUST1001 | 69 | Male | 73925 | 16 | West | 44 |
| CUST1002 | 46 | Male | 67941 | 90 | East | 20 |
| CUST1003 | 32 | Male | 46834 | 60 | West | 17 |
| CUST1004 | 60 | Male | 43047 | 2 | South | 46 |

**Sales Data:**

| **sale\_id** | **customer\_id** | **sale\_amount** | **product\_category** | **sale\_date** |
| --- | --- | --- | --- | --- |
| SALE1000 | CUST1089 | 391 | Clothing | 2022-01-01 |
| SALE1001 | CUST1033 | 549 | Groceries | 2022-01-02 |
| SALE1002 | CUST1006 | 706 | Groceries | 2022-01-03 |
| SALE1003 | CUST1067 | 560 | Electronics | 2022-01-04 |
| SALE1004 | CUST1057 | 376 | Electronics | 2022-01-05 |

**Next Steps for Data Preprocessing:**

1. **Handle Missing Data:**
   * Ensure that there are no missing values in the customer or sales data (e.g., NaN values).
2. **Encode Categorical Variables:**
   * Convert categorical variables like gender, region, and product\_category into numerical representations (using techniques like one-hot encoding or label encoding).
3. **Feature Engineering:**
   * Create new features that could be useful for model training (e.g., total\_spent based on the total of all sales per customer).
4. **Normalization/Scaling:**
   * Normalize continuous features like age, annual\_income, spending\_score, and num\_purchases for use in machine learning algorithms.
5. **Merging Data:**
   * Merge the customer\_df and sales\_df on customer\_id to create a complete view of the customer’s behavior.